
PUBLIC POLICY MASTER'S THESIS

April 2026

**Method in the Madness:
Conceptualisations of User-Generated
Platform Data in the Age of AI**

Aron Thatcher West

Master's Thesis supervised by Prof. Liane Huttner

Second member of the Jury: Prof. Valérie Peugeot

Master of Public Policy
Digital, New Technology and Public Policy

Table of Contents

List of Abbreviations	4
Main Contributions	5
Introduction.....	6
Interdisciplinary State of Knowledge	8
What is AI data collection?	9
What does it mean to “regulate” AI data collection?	9
Analysis of Legal Frameworks for User-Generated Data.....	10
Role of Economics	12
Technical Data Collection Risks	13
Platform Policies	14
Data, Sources, and Methodology	15
Chapter 1: What Data? Conceptualisations of Data in EU Law	17
Relation to a Person.....	17
Availability.....	19
Economic.....	22
Chapter 1 Conclusion.....	26
Chapter 2: What Practices? An Analysis of Platform Policies	27
Relation to a Person.....	28
Platform Policy Alignment.....	28
Legal Regimes, Qualification, and Categories	28
Processing and Obligations	30
Availability.....	35
Platform Policy Alignment.....	35
Legal Regimes, Qualification, and Categories	36
Processing and Obligations	37
Economic.....	40
Chapter 2 Conclusion.....	42
Chapter 3: Towards a Developed Economic Conceptualisation of Data.....	43
Categorisation Processes	44
Value	46
Competition and Unequal Bargaining.....	48
Data Returns and Governance.....	50
Fairness.....	51

Chapter 3 Conclusion	52
General Conclusion.....	53
Policy Recommendations.....	55
Bibliography	57
Annex 1: DSA-designated VLOPs and VLOSEs and their Policies	71

List of Abbreviations

AI	Artificial Intelligence
AIA	EU Artificial Intelligence Act
CDSMD	EU Directive on Copyright in the Digital Single Market
CJEU	Court of Justice of the European Union
CPS	Core Platform Service, defined in the DMA.
DA	EU Data Act
DGA	EU Data Governance Act
DMA	EU Digital Markets Act
DSA	EU Digital Services Act
EC	European Commission
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
EP	European Parliament
EU	European Union
GDPR	EU General Data Protection Regulation
LLM	Large Language Model
NCA	National Competition Authority
NDPA	National Data Protection Authority
NPDR	EU Non-Personal Data Regulation
OCSSP	Online Content-Sharing Service Provider, defined in the CDSMD.
TDM	Text and Data Mining
TFEU	Treaty on the Functioning of the European Union
VLOP	Very Large Online Platform, defined in the DSA.
VLOSE	Very Large Online Search Engine, defined in the DSA

Main Contributions

Online platforms have become widespread in modern life. Holding data of many different types, platforms have significant capacities to direct how data generated by users is processed, where it is used, and the extent to which users are aware. The recent boom in artificial intelligence (AI) development has led to the collection of user-generated data from online platforms for the development and training of AI models becoming commonplace. In this context, both the legislative framework that regulates data use and the platform policies that demonstrate empirical practices have become highly important.

This thesis brings together an overview of European Union (EU) legislation applicable to data with an analysis of platform policies. In the legislation, we identify three principal conceptualisations of data based on, in turn: the relation of data to a person, the availability for processing of data, and the key economic characteristics of data. The last of these, an economic conceptualisation of data, is found to be comparatively under-developed, with central measures for the economic regulation of data resting on alternative conceptualisations and over-simplifications of the economic properties of data. Platform policies examined closely followed the conceptualisations of data identified in the legislation, focusing on personal and availability-focused conceptualisations. Within both, platform policies demonstrated a significant lack of detail concerning the processing of data for AI model development. Meanwhile an economic conceptualisation is similarly neglected.

We therefore examine and propose a strengthened economic conceptualisation of data, identifying how this could benefit both existing legal regimes and, importantly, work to mitigate the economic and market effects of data and its particular characteristics in the context of AI model development using user-generated platform data.

More widely, as data collection for AI model development accelerates and broad changes are being considered to the EU legal framework for data, many of the problems are well-known to legislators, companies, and end users. The desire for fair development and growth at the same time as an emphasis on the protection of privacy and ideas have been framed as fundamentally in tension. This thesis proposes a different way of looking at these well-known problems, aiming to complement rather than complexify. In the world of platforms and AI, this promises benefits for us all.

Introduction

*“POLONIUS, aside:
Though this be madness, yet there is method in 't.”¹*

Communicating with family, friends, colleagues, businesses, schools, politicians, political parties, governments: the list of everyday interactions made possible, and in some cases now only possible, via online platforms is long. Online platforms have played an increasingly significant role in everyday life in Europe for the last two decades. Recent technological developments and changes to lifestyle and communication habits have further entrenched their position in everyday life. From social media to shopping, finding information to navigation, online platforms are now involved in diverse processes across society; it is hard to lead a life without contact with at least some of these platforms. Online platforms rest on digital infrastructure, taking advantage of fast communication over the Internet, readily available processing, and low-cost storage. Using these, they have in some ways become “*digital town squares*”.

Yet these are digital town squares with a twist. Extensive private ownership, business practices based on extensive data processing, and questionable accountability stretch the metaphor to its limits. Data, its processing chief amongst these, is central to the operations and business models of many online platforms, without users necessarily being immediately aware of the consequences. Used for mass service provision, customer analysis, product and advertisement targeting, exchange, and even sale, the digital platform world is founded on data.

Concurrent to the development of online platforms, artificial intelligence (AI) has bloomed into public consciousness since the release of ChatGPT, the first popularised Large Language Model (LLM) AI system, in 2022. AI systems in general, and LLMs in particular, rely upon very large amounts of data. This includes both for the training stage of AI development and for the day-to-day operation of models. AI model training involves calculating statistical weights for tokens of data, establishing patterns of usage in language. Although the relationship between dataset size and model performance is not always strictly linear,² large datasets containing many billions of tokens have enabled LLMs to produce text in many different styles, fields, and languages. Organisations around the world are therefore scrambling to collect large amounts of high-quality data.

The large amounts of data processed by online platforms and their increasing presence in everyday life, combined with the need of AI systems for large amounts of data, creates an important meeting point between platforms and AI development. A series of recent announcements of new AI integrations into online platforms has demonstrated that this overlap has been noted by platform operators.³ Even platforms not directly developing AI

¹ William Shakespeare, *Hamlet*, Folger Shakespeare Library (1601 [2020]), Act 2, Scene 2, Lines 223-4.

² Ilya Shumailov and others, ‘AI Models Collapse When Trained on Recursively Generated Data’ (2024) 631 Nature 755.

³ Anders Giovanni Møller and others, ‘The Impact of Generative AI on Social Media: An Experimental Study’ (2026) 16 Nature Scientific Reports 9376.

models are increasingly integrating AI model training into their processing of data, or collecting data for others to do so.

This thesis seeks to explore the legal bases for the collection of user-generated data from online platforms for the purpose of AI development. The research centres around the principal question: **How is the processing of user-generated platform data conceptualised and regulated under EU law in the context of AI model development?** This is approached through a number of sub-questions, namely: How can user-generated data be conceptualised within the EU legal framework? How do platform policies correspond to the legal framework for the use of data in the context of AI model development? How can the regulation of data processing in the context of AI model development be improved?

Based on the preliminary exploration of the state of knowledge in the section below, it was hypothesised that the current EU regulatory framework for data contains a number of different fragmented conceptualisations of data, a situation which impacts the practices and interpretation of platform policies. An improved alignment between practices and conceptualisations is envisaged by ensuring that the legal frameworks account for all relevant conceptualisations of data.

To approach these questions and hypotheses, this thesis first employs a systemic analysis methodology incorporating the identification of legal regimes and conceptualisations, followed by the identification of current practices, as will be explored further in the methodology section below. The scope of the thesis is limited to the EU legal framework and platforms policies as applicable to EU users.

The thesis first explores the current interdisciplinary state of knowledge and provides detail on the methodology employed. It then examines the conceptualisations of user-generated data in EU law, identifying conceptualisations based on relation to a person, availability of data, and economic properties. Next, it seeks to apply such conceptualisations to current practices of data collection on online platforms through platform policies. Following this, it examines an economic conceptualisation of data and how this can and could interact with and complement existing conceptualisations and legal regimes in the context of data collection for AI model development. Finally, conclusions and recommendations are presented based on the analysis.

It is principally found that although there is significant EU law which utilises conceptualisations of user-generated data based on its relation to a person and availability, the economic conceptualisation of data has been under-theorised and under-developed. This has brought important limitations to the application of the whole legal framework for data, notably in terms of market contestability, personal data protection, and the copyright regime. These limitations have been rendered evident and particularly potent in the context of data collection for AI model training. Meanwhile platform policies narrowly follow such conceptualisations, creating compartmentalised policies which risk not fully representing data collection operations. As with the regulatory framework, information is particularly scant and the terms particularly one-sided when referring to an economic conceptualisation of data. Finally, an economic conceptualisation is found to bring positive benefits to the EU data framework, aiding the application of provisions key to ensuring a fair and contestable market, and more widely to regulating more effectively and efficiently the collection of data for AI training.

Interdisciplinary State of Knowledge

Just as discussions about AI are dominating many political and societal conversations about digital technologies, they are equally in vogue amongst academic and professional circles. Millions of journal articles, books, conferences, lawsuits, regulatory investigations, and discussions over proposed laws have served to sustain this significant discord. Yet key questions remain unanswered.

Even defining AI in the context of regulation can be a tricky task. The EU AI Act uses a lengthy definition based on use cases.⁴ Meanwhile, others have preferred Alan Turing's definition that if a task is labelled as intelligent when done by a human, then it is AI when it is done by a machine.⁵ Evidently this does not categorically solve the definition puzzle, for the subjective idea of "intelligent" tasks completed by humans is not easy to resolve. Meanwhile wider, often non-academic, literature has had a tendency to widely apply the label "AI" to digital technologies that may not strictly merit the term.⁶ Despite such over-labelling, it can be important to initially accept such expansive labels as provided in order to analyse the understanding of the provider; this is the approach adopted for the regulatory frameworks and platform policies analysed herein. There are many different types of AI system, as varied as for text generation, medical applications, and video creation. The predominant focus for the analysis in this thesis is Large Language Models (LLMs). LLMs use very large amounts of text-based data to generate outputs based similarly on text and are widely used across online platforms, thus stand as good candidates for analysis. Yet platform policies frequently do not specify what type of models are in use, and many regulations seek to apply more generally, so references to AI models in the regulation and platform policies explored here should be read generally unless specified otherwise.

Online platforms represent spaces that "*serve at least two different sets of users simultaneously, bringing them together and enabling interactions between them*".⁷ They can take many different forms; the modern web is home to many such platforms.

This section on the interdisciplinary state of knowledge aims to review key areas of the literature relevant to legal conceptualisations of data in the context of the collection of user-generated platform data for AI training. It introduces six relevant areas in the literature which inform the methodology and analysis of this thesis: a definition of AI data collection, what it means to "regulate" AI data collection, how regulatory frameworks for data have been

⁴ Article 3(1) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act, AIA) defines an AI system as: "*a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*".

⁵ Giusella Finocchiaro, 'The Regulation of Artificial Intelligence' (2024) 39 AI & SOCIETY 1961, 1962.

⁶ Leslie Willcocks and Wendy Currie, 'AI Is a Misleading Tale of Hope and Fear' (*LSE Business Review*, 20 March 2026) <<https://blogs.lse.ac.uk/businessreview/2026/03/20/ai-is-a-misleading-tale-of-hope-and-fear/>> accessed 14 April 2026.

⁷ OECD, 'An Introduction to Online Platforms and Their Role in the Digital Transformation' (OECD Publishing 2019) 20.

analysed, how economic ideas have been used in the context of AI and data regulation, the technical risks of AI data collection, and how platform policies have been examined.

What is AI data collection?

“Data” is simply defined by the General Data Protection Regulation (GDPR) as “information”, a definition also adopted and expanded on by the Cambridge English Dictionary,⁸ and our chosen working definition here. User-generated data refers to data created by, or created as a result of, individuals and businesses using online platforms (“users”). Similar terms such as “content” and “material” are understood to be interchangeable with data.⁹ In this sense, “AI data collection” is understood to be information processed for the stated purpose of training an AI model, with the latter understood according to its context. For the purposes of this thesis, such data may be collected for training before or during a model’s operation.

What does it mean to “regulate” AI data collection?

Many scholars have identified issues of data as central to the regulation of AI. These can be divided into regulating data used for training AI models, regulating about the data produced by AI models, and regulating the wider effects of AI on data. It is the first of these which is the focus of this thesis.

The regulation of data collection for AI training raises significant questions concerning the structure and functioning of regulatory systems. Regulation has been typically defined by political scientists as “*the organization and control of economic... activities by means of making, implementing, monitoring, and enforcing of rules*”,¹⁰ used herein.

Regulation can be created in a “top-down” manner, such as by governments, courts, and non- or quasi-non-governmental authorities, or in a “bottom-up” way such as through standards setting and audits. In the context of AI and data, the former has been criticised notably for lacking enforcement means¹¹ and insufficient obligations placed on developers,¹²

⁸ The Cambridge English Dictionary defines data as “*information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer*”. (‘data’ (Cambridge Dictionary) <<https://dictionary.cambridge.org/dictionary/english/data>> accessed 14 April 2026.)

⁹ EU legislation lacks a consistent definition for “content”: the 2022 Digital Services Act defines “*illegal content*” with reference to “*information*”, and the definition of “*terrorist content*” within the 2021 Regulation on the dissemination of terrorist content online refers to “*material*”. Meanwhile, article 2(1) of the 2019 EU Directive on Digital Content and Services defines “*digital content*” as “*data which are produced and supplied in digital form*”.

¹⁰ Walter Mattli and Ngaire Woods, quoted in Cyril Benoît and Matthias Thiemann, ‘Regulation’ in Jon CW Pevehouse and Leonard Seabrooke (eds), *The Oxford Handbook of International Political Economy* (1st edn, Oxford University Press 2021).

¹¹ *ibid.*

¹² Adam Buick, ‘Copyright and AI Training Data—Transparency to the Rescue?’ (2025) 20 *Journal of Intellectual Property Law and Practice* 182.

just as the latter has been criticised for corruption¹³ and complexity leading to lack of understandability.¹⁴

In this way, an analysis of EU data legislation (plus its interpretations) as top-down regulation, and of platform policies as both responses to top-down regulation and sometimes as a distinct form of bottom-up regulation, promises to provide interesting insights into the results and dialogues created through the regulatory process.

Analysis of Legal Frameworks for User-Generated Data

The EU regulatory framework for data has been extensively discussed in the legal literature, frequently focusing on the scope and application of individual legal regimes.

The EU personal data regime, centred on the General Data Protection Regulation (GDPR), is preeminent amongst the individual legal regimes analysed in the context of AI data collection. Scholars have examined its effectiveness through its scope: discussions over the definition of “*personal data*” have argued both that the definition is overly large,¹⁵ or alternatively that its scope is too narrow, with propositions of additional categories of data to incorporate.¹⁶ Meanwhile others have pointed to uncertainties at the edges of definitions, such as the blurry technical distinction between personal and non-personal data to argue that de-personalisation processes are inherently unclear, causing problems for non-personal processing.¹⁷ In applications of AI, the personal data framework has been found by many to be approaching its limits of sufficiency.¹⁸ Scholars have identified structural problems and contradictions within the GDPR’s data processing limitations which clash with the data maximalist approaches of much digital development.¹⁹ Others have taken a more pragmatic approach, proposing smaller “tweaks” to specific measures such as proposing an opt-in to data use.²⁰ Nor are such observations solely recent – literature pointed to flaws in the face of AI before the 2022 LLM-fuelled AI boom.²¹

Scholars have also examined other key EU regimes in the context of AI data collection, including notably the copyright regime, centred on the 2019 Directive on Copyright in the Digital Single Market (CDSMD), and the more economic Digital Markets Act (DMA)-centred legal regime. Text and data mining (TDM), central to current AI data collection exercises and the legality of which is expanded with the CDSMD, has been

¹³ Frederic Heymans and Rob Heyman, ‘Identifying Stakeholder Motivations in Normative AI Governance: A Systematic Literature Review for Research Guidance’ (2024) 6 Data & Policy e58.

¹⁴ Abeba Birhane and others, ‘AI Auditing: The Broken Bus on the Road to AI Accountability’ (arXiv, 25 January 2024); Wei and others (n 13); ‘Big Tech’s Last Minute Attempt to Tame EU Tech Rules’ (*Corporate Europe Observatory*, 23 April 2022).

¹⁵ Nadezhda Purtova, ‘The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law’ (2018) 10 Law, Innovation and Technology 40.

¹⁶ Andreas Häuselmann and others, ‘EU Law and Emotion Data’ (IEEE 2023).

¹⁷ Michele Finck and Frank Pallas, ‘They Who Must Not Be Identified—Distinguishing Personal from Non-Personal Data under the GDPR’ (2020) 10 International Data Privacy Law.

¹⁸ Hannah Ruschemeier, ‘Squaring the Circle’ (*Verfassungsblog*, 7 April 2023).

¹⁹ Hannah Ruschemeier, ‘Generative AI and Data Protection’ (2025) 1 Cambridge Forum on AI: Law and Governance e6.

²⁰ ‘Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World’ (Stanford University Human-Centered Artificial Intelligence 2024).

²¹ Sandra Wachter, ‘Data Protection in the Age of Big Data’ (2019) 2 Nature Electronics 6.

identified as a key area of continuing legal uncertainty.²² More widely, the copyright regime in particular has been identified as upholding a significant inequity between platforms and users, emphasising the former's role as a regulator of content and practices itself, and the lack of control over environment or legal terms granted to the latter.²³ The legal regime around the DMA, often linked with competition law even if it constitutes a different effective legal regime, has divided opinion between those judging it resilient in the face of change²⁴ and those identifying flaws such as its specific focus on tightly-defined Core Platform Services (CPS) meaning that new developments are not explicitly covered by the act, to the potential detriment of contestability and innovation in the market.²⁵

Meanwhile, works surveying the combined effects of several EU legal regimes are also highly present. Surveys of combinations of privacy, intellectual property, competition, product liability, cybersecurity, and consumer law have sought to chart the different evolving legal frameworks.²⁶ Comparative approaches have allowed scholars to identify key areas of overlap and complementary development, with areas of mutual benefit such as between consumer law and data protection law for providing remedies for users.²⁷ Significant tensions between different legislative frameworks have been identified in the context of AI training – memorably termed a “*hodgepodge*”.²⁸ Cross-cutting concerns have notably been discussed regarding inadequate provisions for access to data across the regulatory framework,²⁹ alongside a general lack of transparency and fairness.³⁰ More technical discussions have emphasised the importance of assurances of data quality, reduced discrimination risks, and an exploration of the effects of datasets on innovation within combined legal regimes.³¹

The literature has thus demonstrated a strong compartmentalisation of different legal regimes in the EU, which has led to identified obstacles in application and enforcement. Yet the literature itself has also adopted a rather compartmentalised approach, with the effect of limiting the breadth of analysis. Some have noted the need to widen the lens beyond specific legal regimes to identify and overcome tension between different legal fields, including

²² Rossana Ducato and Alain Strowel, ‘Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out’ (2021) 43 European Intellectual Property Review 322.

²³ João Pedro Quintais, Giovanni De Gregorio and João C Magalhães, ‘How Platforms Govern Users’ Copyright-Protected Content: Exploring the Power of Private Ordering and Its Implications’ (2023) 48 Computer Law & Security Review 105792.

²⁴ Friso Bostoen and Jan Krämer, ‘How Future-Proof Is the DMA? A Case Study Of AI Agents’ (2026) Journal of Competition Law & Economics.

²⁵ Ayse Gizem Yasar and others, ‘AI and the EU Digital Markets Act: Addressing the Risks of Bigness in Generative AI’ (arXiv, 7 July 2023).

²⁶ OECD, ‘Artificial Intelligence, Data and Competition’ [2024] OECD Artificial Intelligence Papers; Claudio Novelli and others, ‘Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity’ (2024) 55 Computer Law & Security Review 106066.

²⁷ Natali Helberger, Frederik Zuiderveen Borgesius and Agustin Reyna, ‘The Perfect Match? A Closer Look at the Relationship Between EU Consumer Law and Data Protection Law’ (2017) 54 Common Market Law Review 1427.

²⁸ Heike Schweitzer and Axel Metzger, ‘Data Access under the Draft Data Act, Competition Law and the DMA: Opening the Data Treasures for Competition and Innovation?’ (2023) 72 GRUR International 337.

²⁹ John Burden and others, ‘Legal Aspects of Access to Human-Generated Data and Other Essential Inputs for AI Training’ (Social Science Research Network, 2 December 2024).

³⁰ Philipp Hacker, Johann Cordes and Janina Rochon, ‘Regulating Gatekeeper Artificial Intelligence and Data: Transparency, Access and Fairness under the Digital Markets Act, the General Data Protection Regulation and Beyond’ (2024) 15 European Journal of Risk Regulation 49.

³¹ Philipp Hacker, ‘A Legal Framework for AI Training Data—from First Principles to the Artificial Intelligence Act’ (2021) 13 Law, Innovation and Technology 257.

through arguments for a generalised classification of EU data law.³² Yet classification exercises can remain rather academic in nature; we propose alternatively a series of groupings centred on the different, sometimes competing, conceptualisations of data in EU law, as will be developed in Chapter 1.

Role of Economics

Scholars have highlighted the extent to which EU digital law has aimed to be informed by and based upon economic principles and analysis. This is the case with the very form of the regulation. Many works have noted the importance of access to data for AI models, with competition law and ex-ante gatekeeper regimes, namely the DMA, notably evoked as tools for ensuring that there is competition in the market for AI model development and training.³³ Although based on many of the same basic aims of ensuring market fairness and contestability, the difference between the DMA and EU competition law is highlighted, with the bar for action on the part of regulators significantly diminished in the name of speed, clarity, and enforceability.³⁴

Within the DMA and competition law frameworks, some have encouraged the development of specific markets for AI training data, arguing that such forms can regulate access using market forces.³⁵ Others have emphasised the role of competition law and DMA principles in facilitating and informing the application of cross-cutting provisions applicable to other legal regimes, such as fairness.³⁶

Meanwhile opinion has varied in economic discussions about data and law over the positioning of data within regulation. As well as much-analysed characteristics of nonrivalry and excludability, data has been shown to have specific economic properties, notably complementarity (where new data processed makes the existing data more valuable) and context-dependence (where the properties and value of data depends on its context), which have made it controversial and difficult to categorise according to traditional models.³⁷ Indeed, some have sought to frame data as an economic good or commodity, including as akin to oil, a common resource, or an essential infrastructure – the first an approach which others have criticised for ignoring the complex ecology around data and its governance.³⁸ Yet the approach has left its mark, with notable enquiries into the extent to which a “value” for data can be identified, identified recently as an undesirable technical possibility but a legal impossibility in the EU.³⁹ Meanwhile ideas of data as a resource or infrastructure have informed discussions on alternate forms of governance for data, included a criticised possible

³² Nine Riis, ‘Shaping the Field of EU Data Law’ (2023) 14 *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*.

³³ OECD, ‘Artificial Intelligence, Data and Competition’ (n 26).

³⁴ Amelia Fletcher and others, ‘The Effective Use of Economics in the EU Digital Markets Act’ (2024) 20 *Journal of Competition Law & Economics* 1.

³⁵ Schweitzer and Metzger (n 28) 341.

³⁶ Hacker, Cordes and Rochon (n 30).

³⁷ Nadya Purtova and Gijs van Maanen, ‘Data as an Economic Good, Data as a Commons, and Data Governance’ (2024) 16 *Law, Innovation and Technology* 1, 15-16.

³⁸ *ibid.*, 39.

³⁹ Gianclaudio Malgieri and Bart Custers, ‘Pricing Privacy – the Right to Know the Value of Your Personal Data’ (2018) 34 *Computer Law & Security Review* 289.

right of ownership for data producers and co-producers,⁴⁰ and forms such as data trusts⁴¹ and data commons,⁴² which are themselves part of a much wider scholarship on common resource governance forms and mechanisms.

The academic discussion demonstrates overarching properties of data alongside a plurality of different economic ideas and possible structures, but lacks widespread agreement on specific forms of regulation. This is evident in that economic ideas have long resided in the legal literature on data regulation, yet often in implicit rather than explicit forms in the legislation.

Technical Data Collection Risks

Both personal data and copyright-related discussions about data have hinged on the extent to which models retain and regurgitate data. This has equally been a concern in more specific discussions such as concerning cybersecurity.

Model developers have claimed that their models merely “*learn*” from training datasets, without retaining the original data.⁴³ Yet this has been contradicted as an explanation by researchers working on the use of copyrighted works to train AI, who have highlighted that vast quantities of training data – such as the entirety of the Harry Potter books⁴⁴ – can be extracted verbatim from current LLMs, gigabytes at a time, despite storage by model developers in the form of mathematical weights.⁴⁵ For personal data, researchers using regular prompts have found LLMs to both fall victim to membership inference and extraction attacks, the former confirming the presence of an individual’s personal data in the training data, and the second extracting verbatim personal data including name, address, and contact details.⁴⁶ Nor is this confined to the most famous individuals who are most present online – social media usernames of a number of individuals were discovered which appeared only in a single document within the training corpus.⁴⁷

The large-scale collection of data of many different types, therefore, for use in AI training is potentially concerning considering how it would seem to be possible to extract it at a later date – albeit potentially in a less accurate, therefore arguably more harmful, form.

⁴⁰ Wolfgang Kerber, ‘Rights on Data: The EU Communication “Building a European Data Economy” from an Economic Perspective’ (University of Marburg 2017) MAGKS Joint Discussion Paper Series in Economics 35–2017.

⁴¹ Sylvie Delacroix and Neil D Lawrence, ‘Bottom-up Data Trusts: Disturbing the “One Size Fits All” Approach to Data Governance’ (2019) 9 International Data Privacy Law 236.

⁴² Stuart Mills, ‘Who Owns the Future? Data Trusts, Data Commons, and the Future of Data Ownership’ (Social Science Research Network, 24 September 2019).

⁴³ ‘How ChatGPT and Our Foundation Models Are Developed’ (*OpenAI*, 30 April 2025) <<https://openai.com/en-GB/policies/how-chatgpt-and-our-foundation-models-are-developed/>> accessed 29 March 2026.

⁴⁴ Ahmed Ahmed and others, ‘Extracting Books from Production Language Models’ (arXiv, 6 January 2026).

⁴⁵ Milad Nasr and others, ‘Scalable Extraction of Training Data from (Production) Language Models’ (November 2023).

⁴⁶ Yupei Liu and others, ‘Evaluating LLM-Based Personal Information Extraction and Countermeasures’ [2025] 34th USENIX Security Symposium.

⁴⁷ Nicholas Carlini and others, ‘Extracting Training Data from Large Language Models’ [2021] 30th USENIX Security Symposium 2633, 2042.

Platform Policies

Finally, the policies produced by platforms themselves have been examined infrequently over recent years, although interest has recently increased in the context of AI training. Certain social media platforms' policies have long been known for their unintelligibility, examined both from a perspective of contract law⁴⁸ and from the perspective of a non-expert user using readability scores.⁴⁹ Such research has argued that the complexity of language and law, combined with the length and number of such policies renders them effectively unreadable. Scholars have equally pointed to the breadth of specific licence terms, especially in the field of copyright law, arguing that users are effectively forced to agree to extremely broad terms to their detriment.⁵⁰ More widely, an analysis of numerous different policies has been a challenge given the sheer number and length of the documents – this was innovatively overcome by a 2023 study which used a RapidMiner text mining machine learning algorithm to pull out and analyse word frequency across documents.⁵¹ This did, however, have the apparent effect of creating a disjointed analysis based on individual words (such as “China”), rather than the legal provisions at hand, which do not always follow the same wording across different platforms and policies.

More recently, in the context of AI, scholars have returned to an analysis of platform policies as a powerful tool for examining platform practices and applications of the legal framework. This is particularly useful for digital affairs where platforms are required to maintain specific policies, yet the exact content is frequently left to their discretion. Quintais et al., conducting a quantitative thematic analysis of the policies of three major platforms (Facebook, Instagram, and YouTube) from an EU user and law perspective,⁵² and King, with a qualitative coding analysis of six AI developers' policies (Amazon, Anthropic, Google, OpenAI, Meta, and Microsoft) with a Californian user and law perspective,⁵³ have both identified significant imbalances of power, information, and accountability between platforms and users. Both studies restricted their legal analysis to specific pieces of legislation, namely the CDSMD and the Digital Services Act (DSA) in Quintais et al.'s case, and the California Consumer Privacy Act in King's case. This inevitably restricted the legal scope of the analysis to EU copyright and California privacy regimes in turn. The articles approached data collection for AI training from opposite standpoints, with the former's focus on user platforms – admittedly all from the Google group, raising questions of how representative the selection was – and the latter's focus on AI developers as the other end of the process. There would therefore appear to be somewhat of a lack of recent wide-scale

⁴⁸ Ellen Wauters, Eva Lievens and Peggy Valcke, 'Towards a Better Protection of Social Media Users: A Legal Perspective on the Terms of Use of Social Networking Sites' (2014) 22 *International Journal of Law and Information Technology* 254.

⁴⁹ Uri Benoliel and Samuel Becher, 'The Duty to Read the Unreadable' (Social Science Research Network, 11 January 2019).

⁵⁰ Miloš Novović, 'User-Generated Content: How Broad Licensing Terms Threaten the Web' *The Responsibilities of Online Service Providers* (Springer International Publishing 2017); Miloš Novović, 'Licensing of User-Generated Content: How Online Platforms Exploit Users' Copyrighted Works' (University of Oslo 2017).

⁵¹ Bisson Christophe, Adele Giron and Gauthier Verin, 'A Comparative Analysis with Machine Learning of Public Data Governance and AI Policies in the European Union, United States, and China' (2023) 13 *Journal of Intelligence Studies in Business* 61.

⁵² Quintais, De Gregorio and Magalhães (n 23).

⁵³ Jennifer King and others, 'User Privacy and Large Language Models: An Analysis of Frontier Developers' Privacy Policies' (2025) 8 *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 1465.

analyses of policies across a range of platforms, an area which nonetheless holds significant potential for examining the effective application of data law frameworks.

Data, Sources, and Methodology

Chapter 1

Chapter 1 seeks to undertake a systematic legal analysis of legislation relevant to user-generated data. EU legal texts, including regulations and directives, were identified using the literature and web searches. In total, 24 texts were analysed, a list of which is provided in the bibliography. These were each examined to identify conceptualisations of data. Law by nature allows for numerous simultaneous categorisations of objects. The process of qualifying allows an object to be categorised, and its category determines the applicable legal regime – or legal regimes, in the case of multiple.

The various categories of data in EU law are compared using a methodology inspired by systems theory. Systems theory was first conceptualised by German philosopher and sociologist Niklas Luhmann. Borrowing from the biological concept of an autopoietic system which can reproduce and recreate itself without outside involvement (said to have operative closure), Luhmann conceptualised modern society using a number of separate systems. Each is made up of a distinct binary code, around which communications create wider meaning and application.⁵⁴ Luhmann recognised that communications between different systems and subsystems are often interlinked as so-called structural couplings, with cognitive openness allowing connections between systems and their environments.⁵⁵ This means that individual elements can belong to multiple systems at once, an important element for legal research. Luhmann identified the legal system as a complex autopoietic system, with a binary code of legal/illegal, and wider communications forming the reasoning that makes an action legal or illegal,⁵⁶ and his theory has been applied to specific areas of law by other scholars.⁵⁷

Inspired by systems theory, this chapter seeks to identify binary codings and to group them. Such groupings are here termed “conceptualisations” of data, each an umbrella to a number of categories and legal regimes.

Chapter 2

Using the systematic framework constructed in Chapter 1, Chapter 2 seeks to examine platform policies in order to identify the conceptualisations of data used and the implications of this on practices concerning the collection of user-generated data for AI training. The

⁵⁴ Richard Nobles and David Schiff, *Observing Law Through Systems Theory* (1st edn, Bloomsbury Publishing Plc 2012) ch 1.

⁵⁵ Stine Piilgaard Porner Nielsen, ‘Introduction: A Special Issue on Niklas Luhmann’s Systems Theory and Law’ (2024) 14 *Oñati Socio-Legal Series* 1206, 7.

⁵⁶ Nobles and Schiff (n 54) ch 1.

⁵⁷ Richard Nobles and David Schiff, ‘Using Systems Theory to Study Legal Pluralism: What Could Be Gained?’ (2012) 46 *Law & Society Review* 265; Katarzyna Gracz and Primavera de Filippi, ‘Regulatory Failure of Copyright Law through the Lenses of Autopoietic Systems Theory’ [2014] *International Journal of Law and Information Technology* 46.

providers of platforms⁵⁸ with the biggest reach in Europe were selected, measured by the highest number of active users. For this, the providers of designated “*Very Large Online Platforms*” (VLOPs) and “*Very Large Online Search Engines*” (VLOSEs) under the DSA⁵⁹ – those with more than 45 million active monthly users in the EU – are used.⁶⁰

Platform policies are not centralised, but rather distributed across providers’ sites. All the platform policies which discuss to at least some extent the processing of data or the development of AI models in relation to the particular service were selected to the best of our ability to identify them. Table 1 in the annex shows the policies and provides links to each. It can already be noted at this stage that the lack of centralisation of such information acts as a serious hinderance of the ability to locate such information about the processing of data, undermining the capacity of an ordinary end user to read and understand the policies and practices of the platform. This joins well-documented concerns over the length and complexity of such documents.⁶¹

Once located, each policy was analysed to identify conceptualisations of data, and to identify areas specific or relevant to the collection of user-generated data for AI development. Platform policies can change frequently – including multiple times a year, representing a further barrier to a common understanding of their terms. The scope of this thesis did not incorporate a longitudinal study over time; policies were collected in February to March 2026, and the analysis therefore represents policies as they stood at this particular moment in time.

Chapter 3

Finally, the third chapter seeks to build on the results of the first two, exploring how an economic conceptualisation of data could function alongside the regulatory frameworks and platform practices identified in the first two chapters.

⁵⁸ For clarity, Bing, a search engine, is also discussed henceforth under the grouping of “platform”.

⁵⁹ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [DSA].

⁶⁰ ‘Supervision of the Designated Very Large Online Platforms and Search Engines under DSA’ (*European Commission*, 1 April 2026) <<https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses>> accessed 10 April 2026.

⁶¹ See, for example, Benoliel and Becher (n 49).

Chapter 1: What Data? Conceptualisations of Data in EU Law

Discussions about data in the digital era have focused extensively on certain conceptualisations of data. The prominence of discussions about data conceptualised by its relation to a person – “*personal data*” and “*non-personal data*” – for instance, has been extensive worldwide. This is particularly the case in EU law, with many regulations and directives using personal data as a key concept – none more so than the GDPR,⁶² the cornerstone of digital law and strategy in Europe which has inspired sister versions across the world. This is not to say that this focus is fallacious. Personal data has been, and remains, immensely significant as a concept, able to be easily understood and referenced by the lay public. But it does not stand alone.

This chapter explores conceptualisations of user-generated data present in EU law. Through a systemic analysis of different legal regimes and categories, it will be shown that data has been conceptualised in different ways in EU law, with regimes and objectives which do not always align closely with each other. It will be demonstrated that a legal and policy focus on particular conceptualisations has at times confused the application of others. Later chapters will explore how different conceptualisations of user-generated data are discussed in platform policies, and more widely how the collection of user-generated data from online platforms for AI training is affected by different conceptualisations of data in the legislative frameworks and platform policies.

In this chapter the principal conceptualisations of data are identified and compared. Some of the regimes examined, particularly the personal and copyright regimes, are enormous in scale; this chapter can only offer only a short introduction to areas relevant to user-generated platform data. **The chapter identifies three principal conceptualisations of user-generated data according to relation to a person, according to availability, and based on economic logic.** This chapter argues that identifying the different conceptualisations of data present in the EU legislation reveals that certain conceptualisations – namely those based on relation to a person and availability of data – have taken precedence over others, notably an economic conceptualisation. This is evident across a number of different legal regimes, where a cross-use of different conceptualisations of data complicates the application of the legal framework.

Relation to a Person

Firstly, by far the most widely used conceptualisation of data is that according to its perceived relationship to a person. The roots of personal data in EU law can be traced back to the European Convention on Human Rights⁶³ which guarantees respect for a private and family life. This principle was legislated more formally in relation to data in the Charter of Fundamental Rights of the European Union, which established the right to the protection of

⁶² Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation, GDPR).

⁶³ Article 8 Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR).

individuals' personal data, the need for a legal basis for processing, and the right of access to such data.⁶⁴ Following this, the so-called "ePrivacy" directive of 2002⁶⁵ was significant for the creation of the EU personal data legal regime, with the restrictions it placed on the processing of data categorised as personal data. The directive lays out requirements for the security and confidentiality of the use of personal data online, including, significantly, preventing tracking without user consent – and therefore hindering the free collection of personal data concerning the users' habits online.⁶⁶

It was, however, the General Data Protection Regulation (GDPR) of 2016 which cemented the relation of data to people as a defining conceptualisation of data. Expanding the ePrivacy directive, the GDPR applies an expansive definition of personal data – "*any information relating to an identified or identifiable natural person*".⁶⁷ Such identification can be direct or indirect, including "*by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*". The GDPR also creates a subcategory of "special categories of personal data", namely data that can be processed to reveal "*racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation*".⁶⁸ This can only be processed with a user's consent or under a small number of exceptions, notably if "*manifestly made public*" by the data subject.⁶⁹

Case law has since clarified that the personal data category includes any data which can be combined to identify an individual, and that anonymisation techniques can only prevent this identification where sufficient other similar data points are present.⁷⁰ Even once direct identifiers (such as names and addresses) have been removed from data, such as following pseudonymisation operations, case law has clarified that data remains personal if a party has "*means reasonably likely*" of being able to identify the data subject using other data.⁷¹ Scholars have pointed to an apparent change in the European Court of Justice (CJEU)'s discussion of the interpretation of identifiability within the category of personal data. Earlier cases⁷² examined both the distinguishability (the ability to single out an individual using the data) and the availability (whether the data could be accessed by an

⁶⁴ Article 8 Charter of Fundamental Rights of the European Union.

⁶⁵ Recital 6 Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications, ePrivacy).

⁶⁶ Article 5(3) ePrivacy.

⁶⁷ Article 4 GDPR.

⁶⁸ Article 9 GDPR.

⁶⁹ Article 9(2)(e) GDPR.

⁷⁰ For a large reading of "personal data", see the inclusion of dynamic IP addresses in *Patrick Breyer v Bundesrepublik Deutschland* [2016] ECJ Case C-582/14; *European Data Protection Supervisor v Single Resolution Board* [2025] ECJ Case C-413/23 P; *Peter Nowak v Data Protection Commissioner* [2017] ECJ Case C-434/16.

⁷¹ *European Data Protection Supervisor v Single Resolution Board* (n 70).

⁷² *Patrick Breyer v Bundesrepublik Deutschland* (n 70).

entity with the power to combine it with other data and identify an individual, different to our discussion of availability below) of data. However, the most recent cases have not considered availability to the same extent, no longer examining whether data *will* be combined, but merely whether it *could* be.⁷³

Yet not all data is personal data. The GDPR's lesser-known sibling, the Non-Personal Data Regulation (NPDR) of 2018⁷⁴ uses the same personal versus non-personal data categorisations to set out rules for the processing of non-personal data within the EU. The NPDR aims to remove barriers internal to the EU affecting the movement of data categorised as "*non-personal*", the definition of which is simply "*data other than personal data*"⁷⁵ as defined in the GDPR.

Therefore, conceptualisations of data which refer to its relationship to a person through its ability to identify individuals can be seen to code data according to a personal/non-personal binary. Qualification using legal categories based on either entails a different legal regime. Article 9 of the GDPR further creates a subcategory of special category personal data, which entails an enhanced specific legal regime.

With the increasing use of AI systems, especially those currently based on LLMs and process very large amounts of data, more and more personal data is likely to be collected as further information is sought for training purposes. Such large caches, combined with the power of LLMs to combine disparate sources of information, risk increasing the possibility of identifying individuals, especially given recent CJEU caselaw apparently lowering the identifiability threshold. This means that data categorised as personal under the GDPR data protection regime increases significantly, potentially incorporating whole AI models.

Availability

EU digital legislation also creates de facto conceptualisations of data. A significant example is data conceptualised by its availability. Although not the subject of specific legislation in the manner of conceptualisations of data according to its relation to a person, understandings of data with reference to its availability are present in a large number of key texts. Availability is used here to refer to whether data can be legally and technically processed by an entity or individual. Data is notably made unavailable under a number of different legal regimes, including notably when it is qualified as copyright-protected, commercial or trade secret, or publicly available.

The EU copyright regime utilises extensively a conceptualisation of data based on availability. The current EU legal framework for intellectual property online is provided principally by the Copyright and Information Society Directive 2001 ("InfoSoc" directive)⁷⁶

⁷³ Sophie Stalla-Bourdillon, 'Identifiability, as a Data Risk: Is a Uniform Approach to Anonymisation About to Emerge in the EU?' (2025) 16 European Journal of Risk Regulation 1456, 1462.

⁷⁴ Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union (NPDR).

⁷⁵ Article 3(1) NPDR.

⁷⁶ Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (InfoSoc).

and the 2019 Directive on Copyright in the Digital Single Market (CDSMD).⁷⁷ The InfoSoc directive creates legal categories of the “works” of “authors” alongside performances, phonograms, films, and broadcasts.⁷⁸ Scholars note that caselaw has since extended the category of an authored work to “all types of works” including original digital content. There is equally a low threshold for originality, where the author must merely have made “*personal creative choices*” in the work.⁷⁹ Once categorised, works are the subject of exclusive rights of use by their authors, with a number of exceptions including solely for transmission or other lawful use.⁸⁰ The CDSMD utilises the same categories, referring to “*copyright-protected works*” to introduce a number of updated exceptions. Particularly notable amongst these for user-generated content is article 4, which establishes an exception for text and data mining, limited only if the rightsholder has “*expressly reserved*” the right.⁸¹ The legal category of a “work” and its surrounding regime can thus be seen to be extensively based on an availability-based conceptualisation of data, viewing it in terms of who has permission to access it, and establishing both a general prohibition on reuse whilst instituting a range of exceptions.

At the same time, the EU copyright law regime also creates categories of entities who process data. Notably, the CDSMD defines an “online content-sharing service provider” (OCSSP) as “*a provider of an information society service of which the main or one of the main purposes is to store and give the public access to a large amount of copyright-protected works or other protected subject matter uploaded by its users, which it organises and promotes for profit-making purposes.*”⁸² Online marketplaces, certain cloud services, open source software platforms, not-for-profit encyclopaedias, and educational or scientific repositories are not included in this definition. However, many of the platforms examined in Chapter 2 are, notably social media and video-sharing platforms, in addition to search engines. Article 17 of the CDSMD establishes that when OCSSPs make copyright-protected works or similar available to the public, they are performing an act of “*communication*” and therefore require a licence from the rightsholder to do so,⁸³ returning once again to an availability-based conceptualisation of data.

Various other legal regimes also utilise a conceptualisation of data according to its (un)availability other than the copyright regime. This is notably the case with commercial and trade secret legal regimes. Trade secrets are created as a legal category and defined by Directive 2016/943 as information which is “*not generally known [within relevant circles]*”, “*has commercial value*” and of which secrecy has been attempted.⁸⁴ The attached legal regime sets out circumstances for the legal and illegal divulgence of trade secrets, once again

⁷⁷ Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (CDSMD).

⁷⁸ Article 2 InfoSoc Directive.

⁷⁹ Quintais, De Gregorio and Magalhães (n 23) 4.

⁸⁰ Article 5 InfoSoc Directive.

⁸¹ Article 4 CDSMD.

⁸² Article 2(6) CDSMD.

⁸³ Article 17 CDSMD.

⁸⁴ Article 2(1) Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure.

framing the regime in terms of availability. Meanwhile, broader mentions of commercial interests are present in several pieces of legislation. The Open Data Directive recognises that when making data public, “*legitimate commercial interests, such as trade secrets, and to intellectual property rights of third parties should be duly taken into account, according to the principle ‘as open as possible, as closed as necessary’*”.⁸⁵ Equally, data are specifically excluded from open access regimes under the 2023 Data Act (DA) for (amongst other reasons), “*commercial confidentiality (including business, professional or company secrets)*”.⁸⁶ The trade secrets legal categories and regime, alongside wider mentions of commercial interests can there be seen to utilise a conceptualisation of data rests on availability.

The opposite end of the scale of availability is also discussed in the legislation. The DA creates the category of “*readily available data*”, defined as “*product data and related service data that a data holder lawfully obtains or can lawfully obtain from the connected product or related service, without disproportionate effort going beyond a simple operation*”.⁸⁷ The recital clarifies that this entails not just technical means, but also the availability of data through design and contractual terms – in other words the legal availability. Meanwhile the Digital Markets Act (DMA) makes reference to “*publicly available*” data.⁸⁸ This term is not defined formally, but examples of specific data types are provided in the same article of the DMA: “*data that is not publicly available shall include any aggregated and non- aggregated data generated by business users that can be inferred from, or collected through, the commercial activities of business users or their customers, including click, search, view and voice data, on the relevant core platform services or on services provided together with, or in support of, the relevant core platform services of the gatekeeper.*” Meanwhile its attached legal regime is merely referred to in the negative in connection with limits on the use of non-publicly available data by gatekeepers – that “[t]he gatekeeper shall not use, in competition with business users, any data that is not publicly available...” (emphasis added).⁸⁹ Both readily available data and publicly available data can therefore be seen as legal categories founded on an availability-based conceptualisation of data.

The attached legal regimes also utilise an availability-based conceptualisation of data, although the interrelation between the regimes is not clear. The DMA’s category of publicly available data appears to be interpreted in much the same way as the DA’s “*readily available*” data with gatekeeper restrictions. These terms, which are not consistent across different texts, appear to function alongside understandings of commercial or trade secrets,

⁸⁵ Recital 28 Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information.

⁸⁶ Article 2(d)(iii) Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act, DA).

⁸⁷ Article 2(17) DA.

⁸⁸ Article 6(2) Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act, DMA).

⁸⁹ Article 6(2) DMA.

and even intellectual property concepts more widely. This lack of consistency creates an obstacle to a clear reading of applicable regimes.

As with personal data, an economic conceptualisation of data is to an extent also present in availability-based legal categories. The DA specifically limits the use of “*readily available data*”, which is specified to be non-personal, when contracted to do so by the user, and further notes that a data holder “*shall not use such data to derive insights about the economic situation, assets and production methods of, or the use by, the user in any other manner that could undermine the commercial position of that user on the markets in which the user is active*”.⁹⁰ Readily available data is therefore recognised to have a potential economic value, including in competition with users. At the same time, the “*commercial value*” of trade secrets is part of their definition, although no further detail is provided on how such value is to be assessed, linking more widely to a lack of an explicit economic conceptualisation explored below.

Data conceptualised according to availability can therefore be seen to be coded according to available/unavailable, with various relevant legal regimes. The EU legal framework utilises extensively an availability-based conceptualisation of data, across copyright, trade secrets, and public legal regimes, and wider linked provisions. Yet the lack of concrete definitions for data availability, with terms referred to usually in the negative, prevents a clear framework from being identified. Furthermore, the qualification of publicly available data also raises a number of questions, notably in terms of what constitutes the public availability of data – a category affected by increasing computing power and ability to search diverse sources of information, in addition to what constitutes an action in competition with the business user. Meanwhile, whilst there is a clear availability-based conceptualisation of data utilised in the regimes explored above, they occasionally borrow notions of data conceptualised according to economic value. Yet these are not fully developed, hindering the categorisation and application of regimes, explored further in the next subsection.

Economic

A third key conceptualisation of data in EU law can be identified which refers to the perceived economic value and properties of data. Such a conceptualisation is present in legal regimes which discuss market value and the power of dominant actors made possible with data. Yet unlike previous conceptualisations, a lack of explicit legal categories or regimes based on a conceptualisation of data as economic has left the variety of economic-related legal texts about data using conceptualisations based on availability and relation to a person.

The cornerstone of EU data legislation dealing in economic terms is the DMA. Based on competition law principles, the act sets down rules affecting competition between digital services including online platforms such as social media platforms and search engines. A key category is that of “*gatekeepers*”, defined as an undertaking which simultaneously has a “*significant impact on the internal market*”, provides a “*core platform service*” used by

⁹⁰ Article 4(13) DA.

business users to reach end users, and enjoys an “*entrenched and durable*” position.⁹¹ The legislation and its recitals discuss at length the economic advantages enjoyed by undertakings which have amassed large amounts of data. These effects are said to be magnified in the digital economy with strong network effects and the nature of online platforms which allow for fast scaling and wide reach. An economic conceptualisation of data, recognising the potential value and market power implications of all data, regardless of other categorisations (i.e. not solely personal or copyrighted data), can be seen as central to achieving the aims of the DMA. Indeed, in the recent context of AI, there is increasing evidence that data is valued for some operations regardless of whether it is personal,⁹² such as with the use of synthetic data to train AI models.

However, rather than utilising an economic conceptualisation of data, the DMA borrows a mixture of different categories from other legal regimes. These utilise a combination of personal and availability-based conceptualisations of data. For example, a specific aim of the DMA was to prevent designated gatekeepers from abusing their position to unfairly compete with business users present on their platforms or in their search engine results. To achieve this, the DMA forbids gatekeepers from using any data which is “*generated or provided*” by business users in the course of their use of the core platform service (CPS) and which is not “*publicly available*” in competition with such users.⁹³ Here, despite the evident intent to qualify with reference to economic value to reduce a gatekeeper’s economic advantage over business users on their CPS expressed by the recitals, the text ultimately uses a framing of availability: the processing of data is restricted for gatekeepers unless it is widely available.

A second central aim of the DMA is to protect individual users of gatekeepers’ core platform services from the adverse effects of market dominance on the part of gatekeepers, and equally to protect business users from the use of such individuals’ data in competition with the business user. However, rather than an inclusive category incorporating data widely, the DMA uses the category “*personal data*” as defined in the GDPR.⁹⁴ The personal data category is then used in the DMA legal regime for the prohibition on gatekeepers from combining datasets or reusing data across their different services. Yet this is at odds with the *ratio legis* of the DMA, created to prevent the cross-use of data across services and thus to prevent the entrenchment of a gatekeeper using its dominant position (such as with a social media platform) to benefit itself (such as on a different platform with an AI system controlled by the same gatekeeper). Use of the personal data category in the prohibition on reuse means that non-personal data can be used with impunity; if a gatekeeper claims to have anonymised the data, it can be freely reused under the DMA. Whilst this may no longer (theoretically) pose a risk for the privacy of users, it does not prevent the gatekeeping platform from gaining on the market using its dominant position. A conceptualisation of data based on economic

⁹¹ Article 3 DMA.

⁹² OECD, ‘OECD Framework for the Classification of AI Systems’, vol 323 (2022) OECD Digital Economy Papers 323 35–39.

⁹³ Article 6 DMA.

⁹⁴ Articles 5-6 DMA.

value rather than borrowing that based on relation to a person would enable a wider application of the prohibition on cross-service use, extended to all data.

Nor is the DMA self-consistent in its conceptualisations. The use of a personal data conceptualisation in the prohibition on data use across gatekeeper services is at odds with the approach taken for business data, for which no limitation to personal data is utilised when prohibiting data combination or reuse.⁹⁵ Although other key business data provisions fail to utilise an economic conceptualisation of data as shown above, the wide scope here stands as a *de facto* recognition that data of many kinds, and not solely personal, can have an economic value. Indeed, under the DMA, gatekeepers are actually required to provide business users with the data of users, including but not limited to personal data, in order to support them, with the users consent and where there is a “*direct connection*” with the business user.⁹⁶ This is a further recognition of the economic value of many different types of data, and more widely stands at odds with the strict data minimisation rules applied by the personal data regime, which is not the desired conceptualisation for this point. Furthermore, the economic framing of data is underlined in the DMA by measures requiring gatekeepers to offer data portability including continuous and real-time access to such data for “*data provided by the end user or generated through the activity of the end user*”.⁹⁷ This measure was introduced in order to support multi-homing and competition between platforms through easy switching, as is explained in the recitals.⁹⁸ It stands also therefore as further recognition of the potential value data can have, including for users.

Data is similarly conceptualised in terms of its availability rather than economic value despite an emphasis on the latter in the Data Governance Act (DGA) of 2022.⁹⁹ The background to the DGA is explained in the recitals in terms of the economic power that data is claimed to bring: “*data-driven innovation will bring enormous benefits to both Union citizens and the economy*”.¹⁰⁰ As with the DMA, there is a strong economic competition focus. The DGA framework for the legal category of “*data intermediation services*” allows users to pool their data and associated data rights,¹⁰¹ intended by legislators to help informed choices on data use. As with the DMA and other provisions of the DGA, there is also an economic aspect to the collective bargaining proposed, which it is hoped will encourage variation – and therefore competition – amongst platform terms.¹⁰² Yet the provisions and associated recitals focus entirely on the “*making available*”¹⁰³ and “*exchange*”¹⁰⁴ of data to such intermediation services and data users. Ignored completely is any discussion of the value of such data, or of the economic effects of different levels of availability. Indeed, no economic basis is provided for the data intermediation services either, a point which has

⁹⁵ Article 6(2) DMA.

⁹⁶ Article 6(10) DMA.

⁹⁷ Article 6(9) DMA.

⁹⁸ Recital 59 DMA.

⁹⁹ Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act, DGA).

¹⁰⁰ Recital 2 DGA.

¹⁰¹ Article 10 DGA.

¹⁰² Recital 31 DGA.

¹⁰³ Article 10 DGA.

¹⁰⁴ Recital 33 DGA.

contributed to the failure of this type of service to become mainstream.¹⁰⁵ In addition, article 16 lays down national arrangements for “*data altruism*”, defined as the “*voluntary sharing of data [...] without seeking or receiving a reward [...] for objectives of general interest*”.¹⁰⁶ This once again underlines the perceived value of data, framed in economic terms – hence it is altruism to grant others access, yet economics is absent save for the prohibition on receiving a reward.

As open structures become increasingly challenged by large-scale commercial usage including to train AI systems, without bringing the market variety and innovation hoped for, the importance of economically based rules of use may become more important. In the meantime, the combination of different conceptualisations of data borrowed across different texts and regimes complicates the application of the very principles enshrined in the legislation. Measures aiming specifically to produce economic effects (reducing gatekeeper market power, encouraging competition and innovation) rest on broad-strokes designations of data to be made available and those to be made unavailable, using personal and availability-based conceptualisations. This makes judging “*availability*” a complex task requiring reference to copyright, commercial secrets, and public data regimes. It also harms the very purpose of the DMA and DGA, namely to regulate gatekeepers and ensure a fair market.

More widely, the tendency to mix conceptualisations of data according to relation to a person and according to availability in the context of regulating economic behaviour leads to a number of overly simplistic assumptions. These include that all data which is available is potentially valuable, a factor in open data policies. Yet simultaneously the legislation assumes that certain data do not hold value, such as non-personal data. By restricting the use of personal data or by making provisions for its pooling or donation, these texts ignore that data generated by a user which is not classed as personal can still hold an economic value. This cross-use of conceptualisations also causes problems within other regimes. For the personal data regime, over-simplified notions of the economic properties of data, notably that it has a fixed value, have led to the apparent acceptance of consent-or-pay frameworks for the use of personal data,¹⁰⁷ introduced in Europe onto Facebook and Instagram in mid-2025. Here, value is assumed to be constant both for data processors and data holders. Yet the potential complementarity of data means that its value can increase as the amount held increases¹⁰⁸ – notable in the case of gatekeepers. At the same time, value is not constant across different users and circumstances, with the impact of such frameworks falling disproportionately on those who lack economic resources and are more likely to consent.¹⁰⁹ Indeed, research by Austrian digital advocacy group NOYB has shown that users consent

¹⁰⁵ See Gabriele Carovano and Michèle Finck, ‘Regulating Data Intermediaries: The Impact of the Data Governance Act on the EU’s Data Economy’ (2023) 50 Computer Law & Security Review 105830.

¹⁰⁶ Article 2(16) DGA.

¹⁰⁷ ‘Opinion 08/2024 on Valid Consent in the Context of Consent or Pay Models Implemented by Large Online Platforms’ (European Data Protection Board 2024).

¹⁰⁸ Purtova and van Maanen (n 37) 16.

¹⁰⁹ Malgieri and Custers (n 39) 19.

within consent-or-pay frameworks at a much higher rate than otherwise,¹¹⁰ raising questions over to what extent such a consent is truly freely-given.

Economic value is therefore a key conceptualisation of data under EU data law, even if it lacks explicit development in the legal texts. As a result of its under-theorisation in law, the texts frequently use conceptualisations of data resting on availability and relation to a person, even in the context of economic aims. This has led to a lack of clarity and provision for certain types of data assumed not to be of value.

Chapter 1 Conclusion

This chapter has sought to demonstrate that user-generated data is conceptualised in multiple different ways in EU data legislation. Three principal conceptualisations are identified: according to its relation to a person, according to its availability, and according to economic properties. Yet despite extensive references to economic reasoning in key texts, there is a lack of an explicit economic coding in this last conceptualisation, with an under-theorisation and under-application of economic categories. Instead, provisions concerning economic value are mixed in the legislation with conceptualisations of data according to relation to a person and availability. Defining data's economic value only in relation to personal data, availability, or other, separate, frameworks risks ignoring and leaving unregulated and undefined important areas of data. Proving our hypothesis for the legal framework, a “patchwork” of different regimes and categories, along with a *mélange* of different conceptualisations, can be seen to confuse and complicate the categorisation of data and application of legal regimes. Furthermore, as data is increasingly collected for AI training with LLMs based on extremely large datasets, the need for an explicit economic coding is likely to become all the more pressing.

The next chapter will examine how the three conceptualisations of data identified – and their applicable legal regimes – are utilised in platform policies, and how they apply specifically to the processing of data in the context of AI systems.

¹¹⁰ NOYB, ‘Pay or Okay: How Companies Make You Pay for Privacy’ (2025).

Chapter 2: What Practices? An Analysis of Platform Policies

Beyond the legislative theory about data, with its categorisations and legal regimes, real-world data practices can be hard to analyse. Data can, after all, be processed without the knowledge of the people concerned, be stored and used at a later date, and lead to effects which may be sporadic and hard to precisely measure.

This chapter aims to analyse the practices of online platforms concerning the processing of user-generated data for AI system development. The platform policies of 20 large providers, the selection of which is explained in the methodology section above, are used to identify platform data processing practices and to examine the legal framework underpinning them. Building on the conceptualisations of data identified in the previous chapter, we aim to compare the conceptual framework with platform policies.

For each conceptualisation, we first examine the alignment between the principal conceptualisations of data in EU law and those used in platform policies, including which legal regimes within the conceptualisation are predominant. We subsequently analyse how user-generated data is qualified according to the categories of the legal regimes applied, including how data for AI training is qualified. Finally, we examine how the terms of the legal regime are applied to the processing of user-generated data in the context of AI training.

We find that user-generated data within platform policies is primarily framed using conceptualisations of data based on their relation to a person and on their availability. Within these conceptualisations, the personal data legal regime of the GDPR and the EU copyright regime are predominant in platform privacy policies and terms of service respectively, although availability-based ideas of public data are also present. Under the GDPR, large amounts of user-generated data are qualified as personal. However, this becomes significantly less clear when referring to data for AI training, with additional conceptualisations of availability frequently introduced. Concerning the application of the legal regime, some platforms failed to define a legal basis for the use of AI. Those that did opted uniformly for the legal basis of legitimate interest, often defining both their interest and the scale of such processing very widely. Under the copyright regime, similarly large amounts of data are qualified as potentially copyright protected. Platforms also apply very widely the terms of the regime, nearly ubiquitously seeking very large licences covering an extensive range of uses of data, including potentially for AI training. Beyond the copyright regime, platforms also sometimes described data as public and appeared to use it to delineate the use of data for AI.

Meanwhile, an economic conceptualisation of data, found to be under-legislated in the previous chapter, is present in many platform policies. Without explicitly applying any specific legal regime, platforms utilise the conceptualisation primarily for asserting that users cannot receive any remuneration for their data, however used. Compared to the extensive detail utilising conceptualisations of data based on its relation to a person or availability, however, the difference is stark with much less discussion of economic value and no categories or legal regime introduced.

A short description of each platform examined is shown in Table 1 in the annex.

Relation to a Person

Platform Policy Alignment

The first chapter identified a common conceptualisation of data according to its perceived relationship to a person. Incorporating the GDPR and NPDR regimes, the conceptualisation deals in terms of privacy and user protection with respect to the capacity to identify individuals.

The platform policies examined use extensively conceptualisations of user-generated data according to its relation to a person. Every platform analysed held a “*privacy policy*” and policies in general which framed their discussion of data for the most part in terms of its relation to a person. Indeed, some platforms went to some length to link corporate values concerning data processing to the personal nature of data. LinkedIn opens its privacy policy by stating that “*Your Privacy Matters*”.¹¹¹ Similar statements are common, including “*we believe strongly in fundamental privacy rights*” (Apple),¹¹² and “[*we aim*] to empower you to make informed decisions when you use X by making sure you understand and have control over the information we collect, how it’s used, and when it’s shared” (X).¹¹³

Within a personal conceptualisation of data, the personal data regime centred on the GDPR was predominant amongst policies examined. Indeed, the terms of privacy policies follow a requirement laid down by the GDPR.¹¹⁴ Platform policies – specifically privacy policies – therefore form the centre of the personal data regime applied to platform processing of data, obliging entities which process data qualified as personal to provide a legal basis, explain the reasons for processing.

Legal Regimes, Qualification, and Categories

As explored in Chapter 1, the personal data regime centred on the GDPR establishes the categories of “*personal data*” and “*special category personal data*”.

Platform policies examined categorise large amounts of user-generated data as personal data, including issuing from searches, orders, viewing, browsing, communication with other users or the platform management, content upload, download, or streaming. TikTok, for example, lists 18 different types of data that it collects and qualifies as personal, ranging from “*User Content*” and “*Technical Information*” to “*Inferred Information*” and “*Content Characteristics and Features*”.¹¹⁵ Importantly, data categorised as personal by platforms is not only data which users are likely to be aware of generating – such as by choosing to write a post or upload a photograph, but also when using the platform without the intention to share data, generating for example technical data. The qualification of such data as personal shows that platforms believe that such data can be used to identify an individual, and therefore attests to a vast collection of identifiable data. Meanwhile four platforms –

¹¹¹ ‘Privacy Policy’ (LinkedIn, 3 November 2025) <<https://www.linkedin.com/legal/privacy-policy>> accessed 15 March 2026.

¹¹² ‘Privacy Policy’ (Apple, 30 July 2025) <<https://www.apple.com/legal/privacy/en-ww/>> accessed 12 March 2026.

¹¹³ ‘Privacy Policy’ (X, 15 January 2026) <<https://x.com/en/privacy>> accessed 15 March 2026.

¹¹⁴ Articles 12-13 GDPR.

¹¹⁵ ‘Privacy Policy (EEA)’ (TikTok, 30 November 2025) <<https://www.tiktok.com/legal/page/eea/privacy-policy/en>> accessed 15 March 2026.

Booking.com, LinkedIn, XNXX, and XVideos – qualify at least some of their data processing as potentially involving special category personal data, all four of which claim would only be inadvertently collected.

Yet such categorisations become rather less clear when pertaining to data processed with AI systems. Only just over half of the platforms studied make explicit reference to AI training within their personal data-related policies. It is unclear whether those that did not do not engage in collecting any data for training AI, claim that data collected does not constitute personal data, or have failed to update their policies. Many of those that mention AI do not refer to data collected for AI training as “*personal data*”, but instead invoke terms such as “*public*” and “*publicly available data*”, avoiding the categorisations of the personal data regime and blending a personal data-based conceptualisation of data with one of availability. Apple, for instance, states that “[w]e do not use our users’ private personal data or user interactions” for training AI.¹¹⁶ Google, Meta, Microsoft, Pinterest, Snapchat, and X similarly claim to use publicly available data. Taken within the personal data regime centred on the GDPR, these statements would suggest notions of “public” personal data and “private” personal data. Yet the regime does not include such categories, nor view personal data being “public” as reason to alter the requirements of its processing. The fact of personal data being “*manifestly made public*” does provide a potential exception to the restriction on processing special category personal data under article 9 of the GDPR. Yet this still does not provide a general legal basis for processing, which is required for all personal data including special category data.

Indeed, some platforms employ terms which sound similar to those within the personal data regime, yet which do not have legal weight. This results in a somewhat confusing suggestion of personal data which is only partially personal. Microsoft, for example, states in the Microsoft Privacy Policy under the heading of “*Does Microsoft use my data to train AI models?*” that “*Microsoft uses data from Bing, MSN, Copilot, and interactions with ads on Microsoft for AI training*”,¹¹⁷ suggesting a use of personal data (as “my” data) from these services. Yet users are assured simultaneously in separate sections that “[i]nformation that may identify you” is excluded and that “[y]our personal interactions with our services are kept private”.

Meanwhile others suggest that they remove or reduce the personal nature of data. Microsoft’s “*Privacy FAQ*” for its AI assistant Copilot states that certain data including searches, interactions, conversation, and voice data are used, but that they have been “*de-identified*”. Under the GDPR legal regime, pseudonymisation can be applied to data in order to reduce the likelihood of identification,¹¹⁸ and anonymisation in order to render it non-personal data.¹¹⁹ “*De-identification*” does not form part of the legal regime, and thus carries no legal weight. Indeed, whilst it would appear to a layperson to refer to removing the capacity to identify (which would create non-personal data), it is treated by the UK ICO as

¹¹⁶ Apple also linked AI data collection to the use of profiling as a part of automated decision-making, stating that it made no attempt to “build profiles about individuals” as a part of AI data collection in its privacy policy.

¹¹⁷ ‘Privacy Statement’ (Microsoft, March 2026) <<https://www.microsoft.com/en-gb/privacy/privacystatement>> accessed 15 March 2026.

¹¹⁸ Article 4(5) GDPR.

¹¹⁹ Recital 26 GDPR.

referring to pseudonymisation, although the body criticises the use of the term.¹²⁰ To users unfamiliar with the specific language of the GDPR (and even for those who are...), this creates a somewhat baffling mixture of terms including “*personal*”, “*private*”, and “*de-identified*”. Such policies appear to seek to reassure users using safe-sounding synonyms, whilst leaving the door open to the processing of personal data for AI.

Processing and Obligations

The processing of data qualified as either personal or special category personal data leads to the application of a number of obligations regarding processing.

Under the personal data regime, all processing of personal data requires a legal basis. Yet whilst demonstrating a vast collection of data, some platform policies examined – notably Amazon, Apple, Microsoft, and Snapchat – do not go into significant granular detail on the nature of processing, retention, or sharing, nor provide a clear legal basis for processing. Amazon, Apple, Snapchat and XNXX’s privacy policies, for instance, do not provide a legal basis for each processing operation, instead merely listing four legal bases and stating that any could apply.¹²¹ Others, most of all Wikipedia, provide distinct and clear explanations of which data are used, and when, how, and under what legal basis.¹²²

The demonstration of a legal basis for processing, already sometimes unclear, was significantly worse concerning processing in the context of AI systems. Despite stating that they collect data for AI training, the policies of numerous platforms failed to clearly state any legal basis at all for such collection. TikTok, for instance, merely states that it collects data “*to develop, test, and make improvements to the Platform, including to develop and improve our technology, such as our machine learning models and algorithms*”,¹²³ not explaining which of the four legal bases listed are in use. Microsoft, which takes a similar approach with an overall privacy policy for its services complemented by certain service-specific policies, does not even list possible legal bases, similarly providing no apparent legal basis for the AI data collection which their own policies shows takes place.¹²⁴

Amongst the platforms which do state a legal basis for AI data collection, every platform examined claimed collection under its “*legitimate interests*”. Such usage is outlined in Article 6(1)(f) of the GDPR, which permits processing where: “*necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.*”¹²⁵ Legitimate interest in the context of AI training has already been the subject of extensive guidance and opinions of data protection authorities. The European Data Protection Board (EDPB) released both guidelines on general processing of data under article 6(1)(f) in

¹²⁰ ‘Introduction to Anonymisation’ (*Information Commissioner’s Office*) <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/introduction-to-anonymisation/>, accessed 13 March 2026.

¹²¹ ‘Privacy Policy’ (*Apple*); ‘Privacy Policy (EEA)’ (*TikTok*).

¹²² ‘Privacy Policy’ (*Wikimedia Foundation*).

¹²³ ‘Privacy Policy (EEA)’ (*TikTok*).

¹²⁴ ‘Privacy Statement’ (*Microsoft*).

¹²⁵ Article 6(1)(f) GDPR.

October 2024,¹²⁶ and an opinion on its use specifically in the context of AI models in December of the same year.¹²⁷ The position of the EDPB has been underlined by the European Data Protection Supervisor (EDPS) in its own guidance produced in October 2025.¹²⁸ The EDPB guidelines set out a three-step test for the use of legitimate interest, which can be summarised as:

- 1) A “*legitimate*” interest of the controller must be being pursued, which is itself lawful, clear and precise, and real and present.
- 2) The processing must be necessary – it could be achieved by no means less restrictive of the rights of data subjects, or with any less data.
- 3) The processing must be balanced against the interests and freedoms of the data subjects, including their reasonable expectations for the use of their data. As a part of this last point, “*mitigating measures*” may be proposed to help balance the analysis of interests.¹²⁹

This three-step test is based on the conditions set out by the CJEU in caselaw,¹³⁰ alongside the requirements defined in article 6(1)(f) of the GDPR.¹³¹ Recital 47 further clarifies that the basis requires a consideration of “*whether a data subject can reasonably expect at the time and in the context of the collection of the personal data that processing for that purpose may take place*”,¹³² supporting the additional mention of reasonable expectations in the EDPB guidance. However, the recognition of mitigating measures is less evidently clear, not appearing in relevant case law,¹³³ but solely on national regulators’ websites.¹³⁴ The concept appears to equate technical capabilities (i.e. to control some element of the processing) with interests, and introduces uncertainty as to how far a data processor is able to weigh the impact of, for example, a settings parameter against an individual’s rights. It would seem logical that any mitigating measures should thus be introduced with a maximum of transparency and clarity over their functioning in order to ensure user awareness. Finally, the legislation makes specific reference to children, yet this is an aspect which is absent from the three-step test. This is important in the context of platforms used by minors, where it is unclear how their expectations may differ from adult users’. Overall, however, the EDPB guidance provides a robust framework for assessing legitimate interests, lacking solely clarity

¹²⁶ ‘Guidelines 1/2024 on Processing of Personal Data Based on Article 6(1)(f) GDPR’ (*European Data Protection Board*, 2024) 1.

¹²⁷ ‘Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models’ (*European Data Protection Board*, 2024).

¹²⁸ ‘Generative AI and the EUDPR: Orientations for Ensuring Data Protection Compliance When Using Generative AI Systems’ (*European Data Protection Supervisor*, 2025).

¹²⁹ *ibid* 7–13.

¹³⁰ Paragraph 106 *Meta Platforms Inc and Others v Bundeskartellamt* [2023] ECJ Case C-252/21: the CJEU identified the three step test as: “*first, the pursuit of a legitimate interest by the data controller or by a third party; second, the need to process personal data for the purposes of the legitimate interests pursued; and third, that the interests or fundamental freedoms and rights of the person concerned by the data protection do not take precedence over the legitimate interest of the controller or of a third party*”.

¹³¹ Article 6(1)(f) GDPR.

¹³² Recital 47 GDPR.

¹³³ For example, paragraph 44 *Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens* [2024] ECJ Case C-621/22.

¹³⁴ For example, ‘L’intérêt légitime : comment fonder un traitement sur cette base légale ?’ (*CNIL*, 2 December 2019) <<https://www.cnil.fr/fr/les-bases-legales/interet-legitime>> accessed 17 April 2026.

with respect to mitigating measures and children. It is used below to examine platform claims to the use of the legitimate interest legal basis.

Of the platforms examined, Booking.com, Google, LinkedIn, Meta, Pinterest, X, WhatsApp, and Zalando all claim legitimate interest for collecting data for AI training. Booking.com, for example, states an interest “*in developing AI systems to reduce our costs, improve the efficiency and quality of our processing, improve our privacy-enhancing technologies and provide better products to our customers.*”¹³⁵ In addition to references to security, LinkedIn includes “[t]o enable economic opportunity and help our Members and customers be more productive and successful”, and “[t]o ensure that LinkedIn’s foundation models are fair, safe, accurate, and effective in the LinkedIn professional context.”¹³⁶ It is therefore clear that some platforms are seeking to use legitimate interest very widely for AI data collection, including with unproven assertions of the ability of AI to provide for the interests of a wide range of beneficiaries. This raises questions over the extent to which the interest can be said to be “*clear and precise, and real and present*” according to the first of the EDPB’s test, which does not yet appear to have been tested in court.

Not only can claims made of the interest of AI training be extensive, but the scale of processing is also often defined largely. Both Google¹³⁷ and Meta, for instance, provide extensive lists of whose interests are served by data collection for AI. Meta seeks to use legitimate interests to “[d]evelop and improve AI at Meta... [i]n our interest, and in the interests of third parties”, a phrase which is repeated in various forms some 15 times within three paragraphs, with a wide range of different data types concerned.¹³⁸ None of the platforms surveyed provided evidence that processing for AI could not be achieved by any other means with less or no personal data, nor of the balancing of users’ interests against their own, once again raising questions over accordance with the EDPB test. Whilst the law and guidelines do not specifically state that such evidence must be available publicly, it would seem to be in the platforms’ interests to inform users, given the importance of users “*reasonable expectations*” to the third element of the test, for users cannot reasonably expect something of which they are unaware.

Despite a lack of information on why a legitimate interest is pursued, a number of platforms offer user settings to control the collection of at least some data, seemingly as a form of mitigating measure under the balancing of interests in the third EDPB test element. The EDPB opinion clarifies that in the context of AI, an “*opt-out*” of data processing may act as a relevant mitigating measure. Such an opt-out is therefore not mandatory, but rather can be put in place at the behest of a platform as part of its balancing of the users’ interests against its own. LinkedIn, Snapchat, X, and Microsoft all offer such opt-outs, apparently permitting users to choose to have their data excluded from processing for AI training,

¹³⁵ ‘Privacy Statement’ (Booking.com, December 2025) <<https://www.booking.com/content/privacy.en-gb.html>> accessed 12 March 2026.

¹³⁶ ‘LinkedIn European Regional Privacy Notice’ (LinkedIn, 3 November 2025) <<https://www.linkedin.com/legal/privacy/eu>> accessed 15 March 2026.

¹³⁷ ‘Privacy Policy’ (Google, 11 December 2025) <<https://policies.google.com/privacy?hl=en-GB&fg=1>> accessed 13 March 2026.

¹³⁸ ‘Privacy Policy’ (Meta, 4 March 2026) <<https://www.facebook.com/privacy/policy/>> accessed 15 March 2026.

frequently using an option in the platform’s settings page.¹³⁹ Importantly, an opt-out mechanism should not be confused with the mandatory right to “*object*” to data processing expressed in article 21 of the GDPR.¹⁴⁰ The right to object permits individuals to request a stop to processing carried out on the basis of public interest (under GDPR article 6(1)(e)) and legitimate interest. Controllers can only continue to process data once an objection has been received if they can demonstrate “*compelling legitimate grounds*” for overriding the rights of the data subject.¹⁴¹ EDPB guidelines clarify that this requires an individual assessment of the impact of processing on that individual,¹⁴² and it therefore cannot be a general policy to process despite objections.

Amongst the use of opt-outs, however, there is a risk of what could be termed “opt-out washing”: the use of an opt-out mechanism which is incomplete or imperfect as a cover for lesser-known activities or to seek to hide other data subject rights. LinkedIn, for example, received significant publicity for its introduction of data collection from European users in November 2025, and the opt-out option introduced by the company was much discussed.¹⁴³ However, a careful analysis of the LinkedIn policies and of the wording on the relevant settings page reveals that the opt-out applies only to the training of “*content-generating*” AI models, and that user data continues to be collected from users regardless of their settings for the development of “*non-content generating*” models, the existence of which is not evident on the platform nor in related media coverage. This fact is not mentioned on the relevant opt-out settings page. Furthermore, the attached “*FAQs*” page discusses the opt-out alongside the statutory right to object, mixing the two different tools. Such practices risk falsely giving users the impression that their data is not being used to train AI, in addition to confusing the optional opt-out with the statutory right to object, and rendering significantly more obscure the process of exercising a data subject’s rights. The risk of misleading users equally works against platforms for the balancing of their interests – and fair expectations – against the desired processing.

In addition to a sometimes-unclear application of the legal test for legitimate interests, some platforms appear to go beyond the provisions of the law by attempting to transfer responsibility for personal data processing onto users. Google’s policy states: “*Please don’t enter confidential information that you wouldn’t want a reviewer to see or Google to use to improve our services, including machine-learning technologies*”.¹⁴⁴ Meanwhile Snapchat’s policy advises users that: “*You should also not share any confidential or sensitive*

¹³⁹ ‘LinkedIn European Regional Privacy Notice’ (*LinkedIn*); ‘Generative AI on Snapchat’ (*Snapchat*, n.d.) <https://help.snapchat.com/hc/en-us/articles/25494876770580-Generative-AI-on-Snapchat> accessed 15 March 2026; ‘Additional Information about Data Processing’ (*X*, n.d.) <<https://help.x.com/en/rules-and-policies/data-processing-legal-bases>> accessed 15 March 2026; ‘Privacy FAQ for Microsoft Copilot’ (*Microsoft*) <<https://support.microsoft.com/en-gb/topic/privacy-faq-for-microsoft-copilot-27b3a435-8dc9-4b55-9a4b-58eeb9647a7f#ID0EFF>> accessed 15 March 2026.

¹⁴⁰ Article 21 GDPR.

¹⁴¹ Article 21(1) GDPR.

¹⁴² ‘Guidelines 1/2024 on Processing of Personal Data Based on Article 6(1)(f) GDPR’ (n 132).

¹⁴³ For example, Wayne Williams, ‘LinkedIn Will Use Your Data to Train Its AI Models Unless You Opt Out’ *TechRadar Pro* (29 October 2025) <<https://www.techradar.com/pro/linkedin-set-to-expand-ai-training-on-user-profiles-heres-how-to-stop-it-using-your-personal-data>>.

¹⁴⁴ ‘Gemini Apps Privacy Hub’ (20 March 2026) <<https://support.google.com/gemini/answer/13594961?hl=en>> accessed 15 March 2026.

information — if you do, it will be used by My AI.”¹⁴⁵ This apparent attempt to transfer responsibility goes beyond the provisions of the personal data legal regime, which centres responsibility for processing, and for identifying a valid legal basis, on the processor. It would also appear disingenuous in multiple ways, being located obscurely in a policy document away from the user interface of either platform’s services, being unclear which services and users it actually applies to, and, most of all, because both platforms’ privacy policies make clear that user-entered information is only a small subset of the personal data collected from users or which is generated automatically with no additional effort on the part of users through their use of platform services. This is especially notable in Snapchat’s case, where the nature of the platform and the use of frequent emojis in the policy make clear that the intended audience is younger people, potentially including under 18-year-olds.

Special Category Personal Data

Aside from the legal basis required for the processing of personal data, data categorised as special category data requires a supplementary legal basis. Of the four platforms who categorised at least some of their data processing as potentially involving special category personal data, only two provided a legal basis for such processing. Under article 9 of the GDPR, special category personal data can only be processed with the explicit consent of the data subject,¹⁴⁶ or under one of a number of narrow exceptions. Both pornographic video-sharing platforms XNXX and XVideos recognise that such data could be processed and claim to use consent as a legal basis.¹⁴⁷

LinkedIn and Booking.com attempt an alternative basis using legitimate interest,¹⁴⁸ although no such basis for processing special category data is established by the GDPR. At the same time, both platforms mention data “*manifestly made public*”. This is an established exception that does allow processing of such data, under GDPR Article 9(2)(e),¹⁴⁹ and touches on conceptualisations of data according to availability as explored in Chapter 1. Yet the extent to which these special category personal data have truly been manifestly made public through publication on either site is unclear. Case law dating from the Regulation 45/2001 on personal data¹⁵⁰ preceding the GDPR established that data made public in one context could not be reused on the basis of it being public knowledge amongst a different public.¹⁵¹ Indeed, the term has been interpreted in a restrictive manner by national data protection authorities.¹⁵²

Whether such processing, which could well be inadvertent on the part of a platform which engages in data collection, therefore would qualify as exempted under this exception

¹⁴⁵ ‘Privacy by Product’ (*Snapchat*, n.d.) <https://values.snap.com/privacy/privacy-by-product> accessed 15 March 2026.

¹⁴⁶ Article 9(2)(a) GDPR.

¹⁴⁷ ‘Privacy Policy’ (*XNXX*, 1 September 2022) <<https://info.xnxx.com/legal/privacy>> accessed 21 March 2026; ‘Privacy Policy’ (*XVideos*, 1 June 2023) <<https://info.xvideos.net/legal/privacy>> accessed 21 March 2026.

¹⁴⁸ ‘LinkedIn European Regional Privacy Notice’ (*LinkedIn*); ‘Privacy Statement’ (*Booking.com*).

¹⁴⁹ Article 9(2)(e) GDPR.

¹⁵⁰ Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data.

¹⁵¹ *Esch-Leonhardt and Others v ECB (ECJ)* [2004] ECJ Case T-320/02.

¹⁵² ‘What Are the Conditions for Processing?’ (*Information Commissioner’s Office*, 18 December 2023) <<https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-are-the-conditions-for-processing/>>.

would therefore have to be examined on a case-by-case basis; caselaw and guidelines suggest that this would likely be interpreted narrowly.

More widely, other platforms who provide no legal basis or discussion of special category personal data processing are also potentially in breach of the requirements. Even limited data collection from such platforms such as for AI development, risks incorporating elements of special category data.

Within the personal data regime, a number of platforms fail to define a legal basis for data processing operations, especially those linked to AI. Of those that do, legitimate interest is the standard selection. However, the interests expressed and the balancing of interests that platforms are mandated to undertake are not always evident. Interests are defined widely or not at all, and the scale of processing subsumed under the interest vast. Matters are further complicated by the occasional presence of an opt-out as a mitigating measure, yet which often turned out to be only limited or confused with the mandatory right to object for users. Finally, special category data appears to be processed by a number of platforms without a legal basis or on the basis of the data being public – itself a dubious claim. The conceptualisation of data according to its relation to a person can thus be seen to be widely used by platforms, but the application of the GDPR personal data regime often unclear in the context of processing for AI development.

Availability

Platform Policy Alignment

The second principal conceptualisation of data identified in the first chapter was based on availability. Availability was understood to mean whether data was legally and technically available for use. Copyright was identified as an important legal regime based on this reasoning, alongside the regimes created around public data, notably present in the context of user-generated data in the DGA and DMA.

A conceptualisation of data based on availability was notable in platform policies, present in some form or other for every platform examined. This conceptualisation was most evident in platform terms, in addition to in wider policies including privacy policies and supplementary policies on AI. Every platform analysed holds some form of “*terms of service*” policy, laying out clauses regarding users’ use of the platform and its services, with the exception of Zalando for which no terms policy could be located. Terms relating to data ownership, rights, and responsibility for data are common. These, alongside the granting of intellectual property licences for use of data, form a central part of the copyright regime applied by platforms.

Aside from the copyright regime, platform policies also feature discussions of user-generated data which refer to public and publicly available data. As will be explored below, these terms approach those used in EU public data legal regimes yet do not appear to actually apply these regimes; rather platforms have co-opted the terms for different purposes.

Although many of the provisions examined in this chapter relate to copyright and public data regimes, terms of service policies contain a broad range of terms. Such terms are laid out in the form of a contract, more evidently so than other platform policies, with

references to “*the agreement between you and [the platform-operating company]*”.¹⁵³ The policy itself forms part of the obligations for providers of intermediary services laid out in the Digital Services Act (DSA). Article 14 of the act obliges providers to publish information on restrictions on the use of their services, including content moderation and details of complaints systems,¹⁵⁴ and to tailor this for minors when appropriate.¹⁵⁵

Legal Regimes, Qualification, and Categories

As demonstrated in Chapter 1, a key constituent of the conceptualisation of data based on its availability is the copyright regime. The EU copyright regime centres on the protection of authorial works, a vaguely defined term in EU law which can include user-generated online platform data. The regime also creates a category for service providers.

Every platform policy analysed applied the copyright regime in its terms of service to a category of data called “*content*”, a word only loosely defined in EU legislation as “*information*”.¹⁵⁶ Some, but not all platform policies introduce their own definitions of content, including Pornhub, for instance, which states: “‘*Content*’ includes the text, software, scripts, graphics, photos, sounds, music, videos, audiovisual combinations, interactive features, textual content, and other materials you may view, upload, publish, submit, make available, display, communicate or post on this Website.” Using this category, platforms emphasise user ownership of data with statements such as “*you own your original content*” (LinkedIn),¹⁵⁷ and “*[y]ou retain ownership of the intellectual property rights [...] in any such content that you create and share*” (Meta).¹⁵⁸ Building on such statements of ownership, responsibility is often similarly transferred to users. Microsoft, for instance, states that “*Your Content remains yours, and you are responsible for it.*”¹⁵⁹ Unlike with the personal data regime and mentions of public data explored below, no specific categorisation within the copyright regime of data used with AI systems is made.

Meanwhile, for platforms themselves, the CDSMD creates the category of “*online content-sharing service provider*” (OCSSP), as explored in Chapter 1. Entities cannot choose this categorisation; it is applied on account of the type of service provided, including notably social media platforms, but excluding online marketplaces. No platform policy examined makes explicit reference to a self-categorisation as an OCSSP. Yet despite this, every platform except Zalando holds terms for the creation of a licence for protected user content, an obligation for OCSSPs.

Chapter 1 also demonstrated the presence of a legal category of “*publicly available*” data which could be applicable to user-generated data. This category forms part of the DMA

¹⁵³ See, for example, ‘Terms of Service (EEA)’ (*TikTok*, August 2025) <<https://www.tiktok.com/legal/page/eea/terms-of-service/en>> accessed 15 March 2026.

¹⁵⁴ Article 14(1) DSA.

¹⁵⁵ Article 14(3) DSA.

¹⁵⁶ For more information on the use of “content” in legislation, see the discussion in the interdisciplinary state of knowledge.

¹⁵⁷ ‘User Agreement’ (*LinkedIn*, 3 November 2025) <<https://www.linkedin.com/legal/user-agreement/>> accessed 16 March 2026.

¹⁵⁸ ‘Terms of Service’ (*Meta*, 4 March 2026) <<https://www.facebook.com/terms/?lang=en>> accessed 15 March 2026.

¹⁵⁹ ‘Services Agreement’ (*Microsoft*, 30 July 2025) <<https://www.microsoft.com/en-gb/servicesagreement/>> accessed 16 March 2026.

prohibition of gatekeeper reuse of business users' data. Yet platform policies' use of "*publicly available*" category does not appear to align with the DMA category, being unrelated to business users. Rather, it appears to be used by platforms without specific legal meaning or attached legal regime.

Unlike with copyright, the de facto public data category is extensively used in the context of processing with AI systems. The policies of Apple,¹⁶⁰ Google,¹⁶¹ and Microsoft¹⁶² all mention "*publicly available*" data in this context: for instance Apple's policy on Applebot model training states that it crawls "*information that is publicly available on the internet*". At the same time, Meta, Snapchat, and X use similar notions of "*public post data*".¹⁶³ The closest possible match of a legal regime to this notion of public availability can be seen as the personal data regime, with its idea of protecting privacy. Yet data being public does not act as a legal basis or justification under the GDPR beyond the specific exception for special category data "*manifestly made public*" examined in the section above. That is to say that categorising data according to its availability and stating that it is public is not a regulatory requirement and does not alter the rules regarding the processing of data. Instead, it appears to be destined to play a more subtle role in reassuring users.

Therefore, whilst the copyright regime can be seen to be applied by platforms with categorisations of user-generated data as potentially protected under copyright law, the same is not the case with the category of publicly available data. Here, the term is used in an apparently different sense by platforms to reassure users about data collection for AI. It should be noted that the DMA and its categorisation of data remain applicable without an explicit categorisation of its data in platform policies, but that policies do not explicitly address the regime or its categories.

Processing and Obligations

Under the EU copyright regime, notably the CDSMD, services categorised as OCSSPs are obliged to obtain a licence from users for making available their content, defined as an "*act of communication*".¹⁶⁴ Such terms are not new; licences have long formed a central part of European digital copyright law.¹⁶⁵

Indeed, all but two of the platforms examined hold policies that express statements on content licensing which are almost word-for-word identical. Pornhub's terms, for instance, state that: (emphasis added)

"by submitting Content to this Website, you hereby grant this Website's operators an unlimited, worldwide, perpetual, non-exclusive, royalty-free, sublicensable, and transferable license to use, reproduce, publish, distribute, broadcast, market, create derivative works of, adapt, translate, publicly display, communicate, or perform,

¹⁶⁰ 'Applebot Model Training and Individual Privacy Rights' (Apple, 23 July 2025) <<https://support.apple.com/en-us/120320>> accessed 12 March 2026.

¹⁶¹ 'Privacy Policy' (Google).

¹⁶² 'Privacy FAQ for Microsoft Copilot' (Microsoft).

¹⁶³ 'Additional Information about Data Processing' (LinkedIn); 'Privacy Policy' (Snapchat); 'Privacy Policy' (X).

¹⁶⁴ Article 17 CDSMD.

¹⁶⁵ See Articles 2-4 Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

***make available or otherwise use all of the Content**, including without limitation for promoting and redistributing part or all of this Website (and derivative works thereof) in any media formats and through any media channels. You also waive to the full extent permitted by law any and all claims against us related to moral rights in the Content. In no circumstances will we be liable to you for any exploitation of any Content that you post.”¹⁶⁶*

Nor is this policy an errant example: all the platforms analysed hold similar terms. The exceptions were Zalando, which does not appear to hold a locatable policy, and Wikipedia, which operates a different licencing framework explored below.

However, although to some extent legally mandated, several scholars have highlighted the vast scale of such provisions and the powers that licences grant to platforms, which may not always be enforceable in reality.¹⁶⁷ Quintais et al.’s analysis of the EU copyright framework notes that the DSA and the CDSMD “[allow] a wide berth for platforms to establish the rules of the game and thereby govern their relationship with users and the content they host”, with users who have “no option but to grant to companies a very broad licence over the texts, videos and images they posted on these services”.¹⁶⁸ More specifically, Novović tracks the presence of copyright licence terms across the major online platforms in 2017 in a couple of articles exploring the wider use of copyrighted data by platforms.¹⁶⁹ Widely-formulated licences and terms grant platforms potentially significant rights to use data for their own purposes. Novović identifies eight key areas where policies apply widely and potentially unclearly: how and when the licence is granted, to whom, which rights are concerned, wider additional terms, general limitations, privacy-related limitations, duration, and standardisation.¹⁷⁰

Amongst the platforms examined (Wikipedia excepted) the licence is frequently granted upon user action, such as uploading information, meaning that users may be unaware of the licence they are granting. Licences are extensive in scale, notably those of AliExpress and LinkedIn, including extending the licence to other, often unclearly defined, affiliate companies.¹⁷¹ Pornhub, Pinterest, Snapchat and TikTok go further still, seeking to grant a licence also to all other users of the platform.¹⁷² The lengths of the licenses proposed varies between specific timescales¹⁷³ to perpetuity.¹⁷⁴ The rights which such terms attempt to establish are also substantial, allowing any and all possible uses of the data by the platform. Even apparent limitations do not clearly restrict platform powers due to their vague wording or lack of restrictive power – such as Apple stating it uses the licence for “internal

¹⁶⁶ ‘Terms of Service’ (Pornhub, 19 November 2025) <<https://www.pornhub.com/information/terms>> accessed 12 March 2026.

¹⁶⁷ Wauters, Lievens and Valcke (n 48).

¹⁶⁸ Quintais, De Gregorio and Magalhães (n 23) 23.

¹⁶⁹ Miloš Novović ‘User-Generated Content: How Broad Licensing Terms Threaten the Web’ (n 50); Novović ‘Licensing of User-Generated Content’ (n 50) ch 2.

¹⁷⁰ Novović ‘Licensing of User-Generated Content’ (n 50) ch 2.

¹⁷¹ ‘Terms of Use’ (AliExpress, 1 March 2025) <https://terms.alicdn.com/legal-agreement/terms/suit_bul_alieexpress/suit_bul_alieexpress202204182115_66077.html> accessed 12 March 2026.

¹⁷² ‘Terms of Service’ (Pinterest, 1 March 2025) <<https://policy.pinterest.com/en-gb/terms-of-service>> accessed 15 March 2026.

¹⁷³ ‘Terms and Conditions’ (Shein, 11 February 2026) <<https://www.shein.co.uk/Terms-and-Conditions-a-399.html>> accessed 15 March 2026.

¹⁷⁴ See ‘Terms of Service’ (Snapchat).

purposes".¹⁷⁵ Across all licences there is a distinct one-sided dimension to the licence. Snapchat's policy continues, for instance, stating that it is not required to "*access, review, screen and delete your content*", despite obtaining a licence to do so.

The licences established by platforms for the use of copyright protected data are therefore often extensive. Whilst it is important to note that such licences do not operate in isolation, but alongside other relevant obligations such as those created by a personal data-based categorisations of data,¹⁷⁶ Novović questions whether licences of such vast scope and applicability are necessary.¹⁷⁷ Indeed, the fact that the size of the recipient, the limitations, and the length vary between platforms would suggest that not all platforms have deemed it necessary to require the maximum scope and applicability in order to operate.

In the meantime, such large-scale licences have the potential to create a jackpot for platform data collection for AI training. Whilst the CDSMD institutes an exception to the requirement for a licence for text and data mining of copyright-protected works, and permits reproduction and extraction of works for that purpose, this is limited by the ability of the rightsholder to opt out.¹⁷⁸ Whereas with platform licences, which users are bound to accept when using the platform, use of data to train AI models could be envisaged even against the wishes of the user, if "*use*" data as expressed by such licences is interpreted broadly.

The sole exception to terms which grant vast licences to platforms amongst the platforms examined is that of Wikipedia. The Wikimedia Foundation requires all content submitted to the online encyclopaedia to be licenced under both the "*Creative Commons Attribution-ShareAlike 4.0 International License*" and the "*GNU Free Documentation License*".¹⁷⁹ At the point of submission of text or non-text media, users are asked to agree to licence it under these two licences, and not to revoke or invalidate such a licence in the future. The two licences form part of an alternative arrangement centred on the free and fair re-use of content, with appropriate attribution, as part of the wider aim of a creation of open-access materials.¹⁸⁰ Whereas conventional platform licences grant wide usage rights to the platforms and their associates, excluding other public uses, creative commons and GNU free licences invert this model, focusing on making data available as widely as possible. This does include potentially for AI, although Wikipedia's policies did not suggest that it collected data for training its own AI systems. The open nature of its licences, however, do mean that all content is available for others to train AI with. Indeed, the Wikimedia Enterprise has expressly leaned into this, providing beta datasets based on structured Wikipedia content in French and English to a number of AI developers.¹⁸¹

¹⁷⁵ 'Apple Media Services Terms and Conditions' (*Apple*).

¹⁷⁶ Dan Neally and Brenda Leong, 'Copyrights and Privacy: What Is the Irrevocable License and Is It Really a Privacy Concern?' (*Future of Privacy Form*, 30 July 2019).

¹⁷⁷ Novović 'Licensing of User-Generated Content' (n 50) 14.

¹⁷⁸ Article 4 CDSMD.

¹⁷⁹ 'Terms of Use' (*Wikimedia Foundation*, 7 June 2023)

<https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use> accessed 21 March 2026.

¹⁸⁰ Digital commons and the Wikipedia model have been extensively analysed elsewhere, and form part of a significant body of literature which is beyond the scope of this chapter to enter into.

¹⁸¹ Including Kaggle, Microsoft, Meta, Amazon, Perplexity, and Mistral AI. (Deborah Mary Sophia, 'Wikipedia Owner Signs on Microsoft, Meta in AI Content Training Deals' *Reuters* (15 January 2026); 'Wikipedia Kaggle Dataset Using Structured Contents Snapshot' (*Wikimedia Enterprise*, 16 April 2025) <<https://enterprise.wikimedia.com/blog/kaggle-dataset/>> accessed 22 March 2026.)

Beyond the copyright regime, data categorised as “*public*” is the potential subject of additional processing, particularly for AI development. X, Google, Microsoft, Pinterest, Snapchat, and Meta all explicitly list public data use for developing AI systems.¹⁸²

The use of self-defined public data by platforms is potentially concerning given that this does not align with any clear legal regime and is not necessarily well understood by users. The categorisation would appear to mean that a child user sharing a post with his or her position on Snapchat’s “*Snap Map*” is processed in the same category of data as a politician’s post on X.

Yet at the same time as use of the unclear de facto category of public data for their collection operations, several platforms sought to institute policies restricting the availability of user-generated data on their platforms, and platform data more widely, to third parties. Apple, for instance, expressly prohibits the use of “*any ‘deep-link’, ‘page-scrape’, ‘robot’, ‘spider’ or other automatic device, program, algorithm or methodology, or any similar or equivalent manual process, to access, acquire, copy or monitor any portion of the Site or any Content*”¹⁸³ in its website terms, including its App Store. Booking and Google have similar prohibitions, with the former explicitly extending to “*AI-powered assistants*”.¹⁸⁴ Once again, such provisions appear outside of any specific legal regime, merely provided as terms in platform policies – which are, after all, forms of contract.

Platforms can therefore be seen to use to a significant extent a conceptualisation of data according to its availability. Large amounts of data are categorised by platforms as copyright protected (under the copyright legal regime) or “*public*” (under no particular legal regime). For copyright protected data, wide copyright terms seek to institute expansive licences which provide platforms with the rights to use very large amounts of data, including potentially for AI training. Meanwhile platforms have also developed their own notions of what constitutes “*public*” and therefore available data, a notion which is already being applied to AI model training. These measures are notably one-sided, leading to the potential for increased processing of user-generated data whilst simultaneously seeking to shut off access to others. Indeed, Quintais et al. find that the relationship between platforms and users in the context of copyright is “*marked by a constant and deep inequity between companies and users as to who controls and benefits from the content circulating in their services*”.¹⁸⁵ The same can equally be said for the whole conceptualisation of data according to availability as used in platform policies.

Economic

The first chapter identified an economic conceptualisation of data as an area under-theorised in law in comparison with personal and availability-based conceptualisations. The

¹⁸² ‘Privacy Policy’ (Meta); ‘Additional Information about Data Processing’ (LinkedIn); ‘Privacy Policy’ (Google); ‘Privacy FAQ for Microsoft Copilot’ (Microsoft); ‘AI at Pinterest’ (Pinterest, n.d.) <https://help.pinterest.com/en-gb/article/ai-at-pinterest> accessed 15 March 2026; ‘Generative AI on Snapchat’ (Snapchat).

¹⁸³ ‘Internet Services Terms and Conditions’ (Apple, 20 November 2009)

<<https://www.apple.com/uk/legal/internet-services/terms/site.html>> accessed 12 March 2026.

¹⁸⁴ ‘Terms and Conditions’ (Booking.com, 15 November 2025) <<https://www.booking.com/content/terms.en-gb.html>> accessed 12 March 2026; ‘Terms of Service’ (Google).

¹⁸⁵ Quintais, De Gregorio and Magalhães (n 23) 23

platform policies examined follow much the same pattern, with no overall conceptualisations of data according to perceived economic value. Notably, the DMA, which Chapter 1 demonstrated sometimes utilises an economic conceptualisation of data yet fails to create any legal categories based on it, is absent in platform policies. In line with the lack of such a conceptualisation, no explicit legal regime based on economic value is introduced or applied in platform policies. Other potentially relevant legal regimes such as the DGA with its data intermediation service framework are also not mentioned.

However, amidst the lack of a legal regime or specific categories based on an economic conceptualisation, platforms do sometimes refer to data with reference to value. Such references are notably one-sided in favour of platforms, particularly those concerning the ability to receive royalties or to charge a fee for data use. For instance, within the intellectual property licence established in platform terms and examined at length above in the section on data conceptualised by availability, all of the platforms with the exception of Booking.com, LinkedIn, and Shein stipulate that such a licence was expressly “*royalty-free*”. LinkedIn and Shein prefer to state that no “*compensation*” would be offered. Snapchat and X further link possible use with the value of data, stating that public content could be used with no compensation to rightsholders. This apparently seeks to establish that users cannot claim any economic value for the processing of their data by the platforms. It would also therefore appear to be a tacit recognition by platforms that user-generated data can have an economic value. Indeed, this is explicitly considered by Shein which reserves the right not only to licence and sub-licence rights to the data to others, but also to “*sell*” the data,¹⁸⁶ and by Pinterest, which states that its terms exclude the possibility of “*any payment or the right to share in any revenue from any monetisation of User Content*”.¹⁸⁷ X even goes so far as to weigh this potential value against a valuation of the use of its services, stating that “*the use of the Services by you is hereby agreed as being sufficient compensation*”.¹⁸⁸

Where platform policies refer to the use of data with AI systems, they are without exception silent on the economic value of data. This is perhaps unsurprising given the ongoing current discussions about compensation for the use of data to train and operate AI systems.¹⁸⁹ The only possible hint is present in the policy of Google, the terms of which include an express prohibition on “*using AI-generated content from our services to develop machine learning models or related AI technology*”.¹⁹⁰ Given Google is unlikely to be concerned about competitors harming third-party model development by feeding AI-generated content in as training material,¹⁹¹ this prohibition could therefore, whilst not explicit, hint at the cost of running current AI models, and the cost that a third-party using Google’s models to train their own models could incur for Google.

In marked contrast therefore to conceptualisations of data based on relation to a person and availability, an economic conceptualisation of data is given little of the light of day in platform policies. With no defined categories or applied legal regimes the

¹⁸⁶ ‘Terms and Conditions’ (Shein).

¹⁸⁷ ‘Terms of Service’ (Pinterest).

¹⁸⁸ ‘Terms of Service’ (X, 15 January 2026) <<https://x.com/en/tos>> accessed 15 March 2026.

¹⁸⁹ Daniel Thomas and Anna Gross, ‘UK to Delay Difficult Decisions on AI Copyright Rules’ *Financial Times* (6 March 2026).

¹⁹⁰ ‘Terms of Service’ (Google).

¹⁹¹ See Ilia Shumailov and others, ‘AI Models Collapse When Trained on Recursively Generated Data’ (2024) 631 *Nature* 755.

conceptualisation is introduced solely in occasional mentions amidst other legal regimes. This has left a series of disjointed and distinctly one-sided remnants, which is notably quiet in the case of AI development. Yet this is a conceptualisation which has vast potential, as will be explored further in the next chapter.

Chapter 2 Conclusion

This chapter has sought to build upon the legislative framework constructed in the first chapter centred on different conceptualisations of user-generated platform data. Through an analysis of the policies of 20 major online platforms in Europe, the conceptualisations used in practice have been compared with those of the legal frameworks. Indeed, there is a strong overlap between the conceptualisations most predominant in the legal texts, notably based on data according to its relation to a person and its availability, and those most present in platform policies. Using these conceptualisations, the personal data and copyright regimes are extensively referred to by platforms, qualifying large amounts of data as personal, copyright-protected, and according to their own notion of public. However, such qualifications become notably less clear in the context of use with AI systems, with a confusing mixture of different legal categories and cross-conceptualisations of data. This was equally the case with the terms of the respective legal regimes, where those related to AI systems were often either absent or starkly low on detail, proving our hypothesis with respect to the impact of a patchwork of legislative categorisations on platform policies. Both personal and availability-based conceptualisations of data can be seen to have, to some extent, failed to properly regulate processing when applied by platforms to processing for AI models.

In comparison, there is a marked lack of economic conceptualisations of user-generated data in platform policies, and a distinct one-sidedness to what is present. Yet an economic conceptualisation of data holds significant potential. This is especially pertinent given the current real-world economic importance of data to AI training and the hence related discussions about the value and need for regulation of such data. The next chapter will examine how an enhanced economic conceptualisation could contribute to simpler and stronger regulation.

Chapter 3: Towards a Developed Economic Conceptualisation of Data

The previous two chapters have sought to explore different conceptualisations of data in the EU legislative framework and in the policies of major online platforms. The first chapter identified a lack of an economic conceptualisation of data within the legislative frameworks created by EU law. Although on the surface several legal regimes including the DMA and the DGA deal in economic terms, it was found that these remained focused principally on conceptualisations of data according to its relation to a person and availability, rather than economic value. Chapter 2 established that platform policies closely resemble in scope the coverage provided by European legislation, and that use of an economic conceptualisation was no exception: the few references made to economic logic were extremely limited and heavily skewed in favour of platforms over their users. This third and final chapter seeks to build upon both prior chapters by exploring how an economic conceptualisation of data could fit within the existing legal framework, how it could affect the relationship between platforms and users, and what such a category could add to the existing framework and legal regimes.

The relative lack of an economic conceptualisation of data is significant given the importance of economic trends to digital platforms, notably concerning data and AI development. Enormous global economies of scale and scope, strong network effects, ecosystem effects, and data feedback loops, as well as at times anti-competitive conduct are common in digital markets.¹⁹² These have had the effect of driving concentration and creating entrenched market positions. AI systems, and particularly the current wave of LLMs, have shown themselves to exacerbate existing economic tendencies in digital markets. The existing domination of the AI model development market by a very small number of companies has already demonstrated strong concentration.¹⁹³ Data is central to AI models, therefore central to the economy surrounding AI development. Indeed, the French *Autorité de la concurrence* noted in its 2024 opinion on generative AI that “[m]ajor digital companies have a huge advantage over their competitors,” through access to data made possible by their operation of services and ecosystems, and more recently with the integration of generative AI tools into these. It further found that major companies “continue to use various strategies to restrict third-party access to their users’ data, by abusing legal rules, such as personal data protection, or security concerns.”¹⁹⁴ In addition to the economic characteristics of digital markets, data also has particular properties which can make a direct transposition of other legal and economic principles difficult. Some, such as nonrivalry and excludability, are well-known.¹⁹⁵ Others, notably context dependence (when the properties of data including its value depend on how, where, and by whom the data is processed)¹⁹⁶ and potential complementarity (when larger data pools bring greater benefits or value),¹⁹⁷ are discussed in the literature but appear to have been somewhat overlooked in the regulatory framework, as explored further in this chapter. The lack of a developed economic conceptualisation of data

¹⁹² Fletcher and others (n 34) 2.

¹⁹³ Hakyung Kim, ‘How Much of the Market Is AI?’ *Financial Times* (18 November 2025).

¹⁹⁴ ‘Generative Artificial Intelligence: The Autorité Issues Its Opinion on the Competitive Functioning of the Sector’ (*Autorité de la concurrence*, 28 June 2024).

¹⁹⁵ Purtova and van Maanen (n 37) 14.

¹⁹⁶ J Mitchell, D Ker and M Leshner, ‘Measuring the Economic Value of Data’ (OECD Publishing 2021) 8;

Purtova and van Maanen (n 37) 15.

¹⁹⁷ Purtova and van Maanen (n 37) 16.

– and connected legal framework and regulation – is therefore particularly significant in the context of AI model development.

This chapter proposes a strengthened economic conceptualisation of data which incorporates three key elements: firstly, a non-binary categorisation (unlike personal or availability conceptualisations where, for example, data is either personal or non-personal), secondly, an appreciation of data's specific economic properties, notably context-dependence and potential complementarity, and thirdly a wider recognition of the value of all data and how this can be relative and variable depending on context, use, and the situation of both holder and processor.

Whereas the literature has explored at length the impact of policies concerning data for AI training on AI developers and the market at large, the impact on everyday users is often overlooked or merely implied in economic analyses.¹⁹⁸ Yet we argue that a stronger economic conceptualisation of data would positively benefit existing European legal regimes which utilise personal and availability-based conceptualisations of data, in addition to render possible a wide range of other potential public policies to the benefit of users, developers, and the market as a whole. This third and final chapter identifies five areas where personal and availability-based conceptualisations of data face challenges in the context of the collection of user-generated data for AI training. These centre on the categorisation processes of data used across different online services, an understanding of the value of data, the identification of and remedies for unequal bargaining and unfair competition, governance and returns from data, and questions of fairness. We argue that an enhanced economic conceptualisation of data, with categories encompassing a deeper economic theoretical background, non-binary categorisation, and a recognition of the variable and relative nature of economic value would permit a strengthening and simplification of the regulatory framework in each of the five areas, whilst pursuing more closely the aims of existing legal regimes.

Categorisation Processes

Firstly, both personal and availability-based conceptualisations can be seen to assume a clear division between different categories of data, and to propose binary categories based on this. For instance, data in the personal data regime is either categorised as personal (and hence the GDPR applies) or non-personal (in which case the NPDR applies). This is similarly the case with the copyright regime and the division between public and private data established by other availability-based conceptualisations. Chapter 2 demonstrated, however, that this has coincided with patchy coverage by platform policies which follow the different categorisations narrowly, creating isolated pockets of information on data use according to a particular regime, and leaving the possibility of significant data processing operations occurring unknown to users. Furthermore, the information provided by policies using narrow categorisations is often further reduced by policies which seek to qualify data minimally with a lack of detail and in an apparently sporadic manner. Indeed, such policies often appear to be racing to qualify data according to the least constraining categories possible, such as by employing passing references to pseudonymisation or “*de-identification*” to suggest a minimised processing of personal data whilst details on the processes remain scant. This situation is worsened by the significant margin left to platforms within the personal data

¹⁹⁸ See, for example Burden and others (n 29).

regime to decide on the appropriate legal basis and to conduct balancing of interests when applicable. However, a patchwork of terms has not reduced their scale: many policies assert extensive processing rights for platforms, such as through broad licence terms granting almost any use of generated data. In this context, the binary categorisations of data, predominantly those provided by personal and availability-based conceptualisations, mean that data which falls outside of restrictive legal regimes exists principally in the negative – data which is neither personal nor copyright, for example.

Yet whilst the divisions are strict between the different categories, with somewhat of a void left for data which falls “between the cracks”, at the same time both platform policies and the legislation mix legal categories. For example, as noted in Chapter 1, the DMA, explicitly created in response to the economic tendencies of digital services, forbids designated gatekeepers from combining or cross-using “*personal data*” across different services provided by the gatekeeper,¹⁹⁹ and also bans the use of business data – “*any data that is not publicly available that is generated or provided by [...] business users*” – in competition with business users.²⁰⁰ The limitation to personal and business user data respectively has the effect of worsening the patchwork of categorisations of data and letting potentially significant data processing continue unabated. Indeed, the references to personal and public data appear out of place given the DMA’s aim of ensuring contestability in digital markets and protecting business and other end users against unfair practices,²⁰¹ and therefore economic focus. Echoing the legislation, many of the platform policies examined also mixed personal and availability-based conceptualisations of data to justify processing despite gatekeeper limits. Chapter 2 illustrated that certain data are currently treated as freely available in platform policies for AI training under both personal and availability-based conceptualisations on account of the configuration of the user’s account – namely if it is set to “*public*”. This patchwork of categorisations, therefore, neither captures the complete processing of data by platforms nor appears to control it.

The explanation for this mixture of conceptualisations of data, and therefore of different legal regimes, would appear to lie with an overly tight focus on individual issues of the day. At the time of the legislation’s development, before the wave of LLM AI systems, it would seem that the main preoccupation was one of personalised advertising – it is in this context that the DMA seeks to limit the re-processing of personal data accumulated through use of platform services, and ensure the availability of “*less personalised but equivalent alternatives*”.²⁰² This is underlined by the inclusion of “*online advertising services*” as a standalone CPS.²⁰³ Yet with the rise of data collection for AI training, it is no longer the collection of solely personal or business data by a gatekeeper which risks distorting the market. As a result, key provisions on the reuse of data across core platform services, portability of data, and the identification of CPS currently lack clarity and power due to a lack of a widened incorporation of data based on its economic value.

Working alongside personal and availability-based conceptualisations of data and their associated legal regimes, a strengthened economic conceptualisation of data offers the

¹⁹⁹ Article 5(2)(b) and (c) DMA.

²⁰⁰ Article 6(2) DMA.

²⁰¹ Recital 2 DMA.

²⁰² Recital 36 DMA.

²⁰³ Article 2(2)(j) DMA.

opportunity to overcome narrow compartmentalisation of data through a non-binary and wider categorisation system. Data does not have to be “economically valuable” or “not valuable”, for example, but can exist in a number of shades. This would function alongside its categorisation in other regimes, therefore without loosening personal or availability-based protections. Meanwhile a wider category incorporating all relevant, potentially valuable, data could help reduce some of this confusion by requiring transparency of how all user-generated data is processed by the platform, including for AI development. Such a regime could help to overcome the increasingly significant lack of information on AI training which leaves users in the dark over the complete use of data.

Value

An economic conceptualisation of data would also clarify and harmonise the legal framework with reference to the value of data. Currently, the legislation makes different simultaneous assumptions about data’s value. Although the DMA recognises in principal that data in general can be valuable (the recitals to the act point to the “*benefits from data*” which can limit contestability amongst digital services²⁰⁴ and frame data as a resource gathered from users of digital services designated as gatekeepers), the terms of the act assume that data is solely valuable for the gatekeeper if it is personal or related to a business user, and hence put in place restrictions on the re-use of such data. At the same time, the DA effectively assumes that all data originating from connected physical devices which is not personal is valuable,²⁰⁵ the reuse of which should therefore be maximised with the aim of encouraging innovation.²⁰⁶ In light of this, the DA seeks to establish a horizontal right to portability regardless of gatekeeper status, across all service providers, aiming for “*an optimal allocation of data for the benefit of society*”.²⁰⁷

Meanwhile traditional copyright and intellectual property law rest on the principle that original ideas are valuable, hence seeks to apply protections for such ideas in order to maintain the incentive to produce them.²⁰⁸ Article 17 of the CDSMD explicitly extends this regime and understanding to data uploaded by users.²⁰⁹ Yet the complex relative value of data and its wide range of use cases is not necessarily well represented by the standard copyright focus on the fixed value of established, marketed works such as books, films, and music. Nor is it necessarily the “*idea*” expressed in the data which has economic value, but rather the data itself. Applying standard valuation logic, the European copyright regime has required platforms to request licences for the distribution of users’ data, qualified as a “*communication*” activity. These requirements appear to have been envisaged by legislators as a means of preventing the upload and sharing of specific copyrighted material by users who do not have the rights to do so, such as the illegal uploading of films, for which users, not being the rightsholder, are not able to legally agree to such licences.²¹⁰ Yet the implicit assumption that only certain works hold value and that data can simply be regarded as a work, and which does not appear to have widely prevented the uploading of individuals’

²⁰⁴ Recital 32 DMA.

²⁰⁵ Recital 2 DA.

²⁰⁶ Kerber (n 40) 3.

²⁰⁷ Recital 2 DA.

²⁰⁸ Kerber (n 40) 4.

²⁰⁹ Article 17 CDSMD.

²¹⁰ See Recitals 60 and 61 CDSMD.

copyrighted content,²¹¹ has meant that some data have been neglected. Users have ended up with reduced control of the data they generate and have not been able to enjoy the benefits normally granted by copyright protection such as a return for use. Furthermore, the legal terms anticipated by such legislation are now frequently complemented by platforms with additional, more stringent terms, to which users are required to agree in order to use the platform, and which frequently include terms stating that users can receive no return for the use of their data. This is not, therefore, a problem with the copyright regime, but rather a limit of its capacity to handle data which has different and variable economic characteristics.

The understanding of the value of data in the legislation is thus somewhat disjointed. Yet as LLMs have made clear, the combination of any data, whether categorised as personal, copyright, or anything else, can be valuable for purposes not originally envisaged by legislators, such as training AI models. In this way, it is not the relation to a person, its origin from business users, or the copyrighted “*work*” which makes the data potentially valuable, but its existence as data. In this context, an economic conceptualisation of data permits a wide category of data – or several – which recognise that all data has a potential value. Existing regimes, such as those created by the DMA and which currently use an awkward combination of different conceptualisations, could be clarified to include a category of user-generated data recognising its potential value, and taking into account non-regime specific factors such as its complementarity. This would not alter the overall direction of the text, instead arguably rendering it more in line with the objectives expressed in the recitals. More widely, an economic conceptualisation could aid the production of clearly referenced standards for datasets and exchanges of data. Scholars have noted the need for a legal framework to ensure consistent standards for data exchanges,²¹² which would notably be useful to developers for gauging the quality of data received when exchanging and marketing datasets.

Furthermore, in the context of AI training an economic conceptualisation of data would aid the formal recognition and incorporation into specific regimes, notably the DMA, of the variable value of data according to the services for which it is processed. The DMA empowers the European Commission to designate gatekeepers with reference to their platform services, the categories of which incorporate aspects of economics.²¹³ Unlike personalised advertising services, the DMA does not currently regard AI as a distinct CPS. This means that restrictions on the transfer of data between AI services and other service provided by a gatekeeper, such as social networking, do not clearly apply. Indeed, many platform AI services are currently regarded by platforms as part of the main CPS, such as LinkedIn’s AI services which LinkedIn categorises as part of its social networking platform service.²¹⁴ In addition, nothing in the DMA would appear to prevent platforms from using an AI trained as a part of a CPS with a different service, or selling the trained model to a third party. There exists therefore a significant disparity between the identification of personalised

²¹¹ Martin Wolf, ‘Playing “Whack-a-Mole” with Meta over My Fraudulent Avatars’ *Financial Times* (25 April 2025).

²¹² Hacker (n 31) 283; OECD, ‘Artificial Intelligence, Data and Competition’ (n 26) 31.

²¹³ Fletcher and others (n 34) 10.

²¹⁴ ‘Microsoft DMA Compliance Report 2026: Annex 11 – LinkedIn (Online Social Networking Service)’ (2026) 13 https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/mscle/documents/presentations/Microsoft_DMA_Compliance_Report%e2%80%933LinkedIn_Section_2_Annex.pdf accessed 1 April 2026.

advertising as a separate service and the non-identification of AI services. A strengthened economic conceptualisation of data would further inform and encourage a designation of AI services, which gather and process huge amounts of data, as a category of CPS. This move has broad academic support²¹⁵ and has formed part of the public consultation on the DMA.²¹⁶ Looking forward, taking a broader understanding of legal categories of data with reference to their economic value, itself recognised as variable, promises to allow greater adaptability in the face of new technologies such as AI.

Competition and Unequal Bargaining

An economic conceptualisation of data would further aid the Commission's express aim of tackling unequal bargaining and its effects on digital markets.²¹⁷ The presence and effects of dominance on a market is an area which has long been explicitly centred on economic analysis. The competition law regime has been central to this. Unlike specific legislative acts such as the DMA which seek to establish across-the-board ex-ante rules for maintaining a competitive market, competition law acts on an ex-post case-by-case basis. At its core, EU competition law seeks to control agreements that harm competition, control the conduct of dominant firms, and control mergers and acquisitions that could impact the market.²¹⁸ Article 102 of the Treaty on the Functioning of the European Union (TFEU) prohibits the abuse of a dominant position within the market.²¹⁹ The article sets a high bar for the identification of abuse, requiring simultaneously that an undertaking holds a dominant position on the relevant market, that such a position has been abused namely by refusing access to other undertakings, and that there is no justification for such refusal.²²⁰ Importantly given the complementarity of data, access to data has been recognised as relevant for assessing the position of undertakings in a market,²²¹ notably in the *Google Shopping* and *Google Android* cases.²²² Furthermore, the use of data across different services has also been the subject of a fine issued to Facebook by the German Bundeskartellamt in 2019.²²³ However, EU competition law cases have not recognised to date how data amassed in the context of one service in a particular market (such as that social media platforms) can affect a services in a different market (such as that of AI training). Indeed, the identification of the “*relevant market*” is often the centrepiece of such actions, with dominance largely understood in Europe to be expected if an undertaking controls more than 50% of such a market.²²⁴ Yet identification of the relevant market can be challenging in digital markets involving data where the scale and presence of alternatives can be unclear.²²⁵ An economic conceptualisation of data could aid this situation by formalising a recognition of the roles that data can play on the market and the importance of complementarity and the context to its

²¹⁵ Yasar and others (n 25); Bostoen and Krämer (n 24) 38.

²¹⁶ Megan Kirkwood, ‘Will the EU Designate AI Under the Digital Markets Act?’ (*Tech Policy Press*, 26 September 2025).

²¹⁷ European Commission Data Economy Communication, quoted in Kerber (n 40) 4.

²¹⁸ Burden and others (n 29) 22.

²¹⁹ Article 102 Consolidated version of the Treaty on the Functioning of the European Union (TFEU).

²²⁰ Burden and others (n 29) 25.

²²¹ Schweitzer and Metzger (n 28) 341.

²²² Case T-612/17 *Google Shopping* ECLI:EU:T:2021:763; AT.40099 *Google Android* Commission Decision [2018] C(2018) 4761 final.

²²³ *Bundeskartellamt, Facebook* (Case B6-22/16) [2019] Decision of 6 February 2019.

²²⁴ Burden and others (n 29) 23.

²²⁵ *ibid.*

value. For example, framing data as an economic good in its own right, especially in the context of AI training, leads to the recognition of a market for data which could then feed into competition law frameworks.²²⁶ Given the specific categorisations of data remain under discussion,²²⁷ a widened economic conceptualisation should retain sufficient flexibility to incorporate alternative categorisations of data including, for instance, as a common.

Even with an enhanced economic conceptualisation of data, competition law cases would likely remain rare given the expansive scale, extensive time requirements, and therefore expensive nature of legal challenges for petitioners. The DMA was expressly created to provide faster ex-ante remedies for the digital market and to move beyond case-by-case enforcement. It aims to asymmetrically target platform-operating enterprises which hold a dominant position on the market and to act even in cases where the high bar of Article 102 of the TFEU may not be reached.²²⁸ In the same manner as with the competition law regime, a strengthened economic conceptualisation of data would permit a contextual recognition of how amassed data can be complementary regardless of legal category – for instance personal data can be combined with copyright protected data – and furthermore that its value is context-dependent, notably on the position of its processor. These factors can lead to data controllers having a “*significant impact on the internal market*”, and therefore falling to be designated a gatekeeper.²²⁹ They could even allow for an expansion beyond the DMA framework based on the designation of specific core platform services, an approach which relies on fixed categories of platform service and which has already shown itself to require regular modification such as with recent discussions on including AI model development. Instead, an approach based on the resources a gatekeeper holds – its situation – could be more heavily emphasised.

Beyond competition between platforms, key texts also aim to tackle unequal bargaining between platforms and users. Central to this is Article 6(9) of the DMA which requires portability of “*data provided by the end user or generated through the activity of the end user*” free of charge and in continuous, real time.²³⁰ The GDPR and the DA hold similar provisions.²³¹ This provides users with potential bargaining power by reducing lock-in effects which effectively force agreement with platform terms. Indeed, portability is explicitly justified in the recitals of the DMA as being intended “[t]o ensure that gatekeepers do not undermine the contestability of core platform services, or the innovation potential of the dynamic digital sector, by restricting switching or multi-homing.”²³² Framed in this way, data portability inherently utilises ideas of the variable, context-dependent value of data – made more valuable to both parties through regulation requiring its portability – to increase user empowerment. At the same time, data intermediation services in the DGA also serve as tools

²²⁶ Whilst the core of EU competition law is defined centrally in articles 101 and 102 TFEU, National Competition Authorities (NCAs) have the right under Article 1(6) DMA to impose their own national competition rules as long as these do not contradict Commission decisions or rely on the gatekeeper status to impose additional conditions. NCAs equally retain the right to apply EU-wide competition law rules on their territory under Council Regulation (EC) No 1/2003 of 16 December 2002 on the implementation of the rules on competition laid down in Articles 81 and 82 of the Treaty.

²²⁷ Purtova and van Maanen (n 37) 2.

²²⁸ Recital 5 DMA.

²²⁹ Article 3(1) DMA.

²³⁰ Article 6(9) DMA.

²³¹ Article 20 GDPR; Article 30 DA.

²³² Recital 59 DMA.

to increase users' bargaining powers through creating collectives, described in the recitals as intended to encourage platforms to compete for users by proposing alternative terms of service.²³³ In this way, portability and data intermediation services stand as examples already utilising an economic conceptualisation of data, even if implicit in the GDPR and DA.²³⁴ It is notable that the flaws identified with portability come not from its economic role, but from other conceptualisations; scholars have noted that access to data within the data protection regime is limited and personal data is used as a barrier to portability.²³⁵

In this way, a strengthened economic conceptualisation of data could assist in applying competition law principles, whether within the competition law regime or more widely in ex-ante frameworks such as the DMA. The conceptualisation would align closely with and complement the existing provisions for data portability found in a number of texts. In the longer term, if utilised to a greater extent, measures such as the portability of data, alongside more equal bargaining and competitive markets assured by competition law could encourage the development of alternative digital models, such as those which propose with different data processing operations or which commit to more restrictive uses of data.

Data Returns and Governance

A reinforced economic conceptualisation of data additionally offers the possibility of a more nuanced discussion around the governance of data and potential returns for users for its use. Such principles currently exist in EU legal regimes, notably with discussions of data altruism, data intermediation services (also known as data trusts), and remuneration. These regimes often use characterisations of data as economically valuable, but mix different conceptualisations of data. For instance, applications offering to pay users for the use of their personal data,²³⁶ stand as an example which combines personal and economically based conceptualisations of data. However, these offers also have well-documented flaws, notably in terms of the disproportionate impact on those with low incomes²³⁷ and also pose difficulties with later withdrawing consent, a particularly salient issue in the context of AI training where data cannot be easily removed from training corpuses or trained models. In many ways, these flaws can be seen to originate from the creation of false equivalencies between different conceptualisations of data combined with fixed notions of its value, notably here assuming that data being personal means that it is valuable economically and not the inverse. As explored above, this may apply to certain uses of data – notably for personalised advertising – but cannot be assumed to always be the case. AI in particular threatens to overturn such assumptions. Similar false equivalencies can be seen in “*consent-or-pay*” frameworks. Once again this can be seen to have originated with an idea that the data, predominantly conceptualised as personal in such discussions, has an innate value and therefore that a one-size-fits-all price for the non-processing of such data can be put into place. Simplistic models similarly applied to AI training risk widening these issues, and creating a situation where the data of those with sufficient means is not processed for

²³³ Recital 31 DGA.

²³⁴ Schweitzer and Metzger (n 28) 346.

²³⁵ Schweitzer and Metzger (n 28) 338–9.

²³⁶ See, for example, TADATA: Alba Ilari, ‘Le cas TADATA’ (*DL4T - Université Cote d’Azur*)

<<https://droit.univ-cotedazur.fr/dl4t/le-cas-tadata>> accessed 10 April 2026.

²³⁷ Malgieri and Custers (n 39) 19.

training, whereas those with lesser means are subject to far greater processing, with all the ramifications that it may engender.

Meanwhile the EU copyright regime also mixes different conceptualisations of data and makes assumptions about the comparative value of certain types of data over others. Technical availability-based measures have been imposed under the regime on large platforms including systems to provide remuneration or to detect and prevent the re-upload of copyrighted content if requested by the rightsholder.²³⁸ Yet these obligations have favoured large copyright-owning enterprises such as major music labels whilst providing no provision for independent artists or even everyday users' content. At the same time, such remuneration remains far from a rigorously conceptualised provision, with long-standing questions over the proportionality of remuneration offered²³⁹ and more recent concerns over remuneration for the use of data with AI models.²⁴⁰

A strengthened economic conceptualisation of data, whilst not a panacea, would permit a more nuanced regulatory model of remuneration potentially based on more precise categories of data, the use of the data, and on the circumstances of holders and subjects, recognising data's value as relative and seeking to move beyond one-size-fits-all models of value. This could help avoid either an overly large offer of remuneration which could harm users,²⁴¹ or an overly constrictive one which effectively gatekeeps who can make money from content creation and who cannot.

Fairness

Finally, an economic conceptualisation of data could tackle issues of fairness, also identified as a priority by the Commission. This is in line with existing legislative aims, notably with the personal data regime which is in part a recognition that market forces alone cannot fully protect individuals' fundamental right to a private life and therefore to the protection of data which identifies them.²⁴² A legislative framework which protects users of online platforms is important given their effective status as de facto "digital town squares" which are now often intrinsic parts of everyday civic life, alongside the wide-scale domination of such spaces by a small number of dominant actors which can dictate the law and rules of such spaces. In this context, fairness can be seen to hinge on all three principal conceptualisations of data: relation to a person on account of the need to protect users' private lives, availability on account of the importance of such spaces and data to users and the wider community, and economic on account of the significant market forces at work in such spaces. All three can be seen as simultaneously important to the fair functioning of online platforms.

²³⁸ Article 17(6) CDSMD.

²³⁹ Janek Meyn and others, 'Consequences of Platforms' Remuneration Models for Digital Content: Initial Evidence and a Research Agenda for Streaming Services' (2023) 51 *Journal of the Academy of Marketing Science* 114, 115.

²⁴⁰ Daniel Thomas, 'UK Creative Sector Seeks Copyright Protection from AI Companies' *Financial Times* (16 December 2024).

²⁴¹ Malgieri and Custers (n 39).

²⁴² Recital 7 GDPR.

Chapter 3 Conclusion

An economic conceptualisation of data was identified as both an under-theorised element of the legislation in Chapter 1 and as present in a one-sided manner in platform policies in Chapter 2. This chapter has sought to establish what an economic conceptualisation of data could mean, and how it could align with predominant conceptualisations according to relation to a person and availability, as well as the related legal regimes and categories.

We first establish key elements of a strengthened economic conceptualisation of data, notably a non-binary framework, a stronger economic theoretic background particularly with reference to complementarity and context dependence, and a widened understanding of the variable value of all data. We argue that a strengthened economic conceptualisation of data would not only provide a rigorous means to regulate the collection of user-generated data for AI development, but would also align well with other conceptualisations of data, strengthening and even simplifying the application of existing legal regimes. A non-binary categorisation system would allow for a clearer and more comprehensive application of legal regimes, moving beyond a patchwork of isolated categories. A greater recognition of the economic properties, notably complementarity and context-dependence, of data would better capture the effects of large data collection and processing, helping to prevent market concentration. Meanwhile an understanding of the context-dependent, variable value of all data would aid even regulatory coverage whilst allowing for greater nuance in discussions over the governance of and returns from data, whilst reinforcing action for greater fairness in the world of online platforms.

An economic conceptualisation of data, however, is not a panacea and will not resolve all problems. An over-reliance on economics can overshadow the importance of protecting users as human beings, rather than market entities. Yet amidst increased processing of data for AI model development, where economic forces have defined the functioning of platforms as shared online spaces, we argue that an economic conceptualisation of data is increasingly important, alongside personal and availability-based conceptualisations.

Furthermore, an economic conceptualisation of data allows for a wide number of specific public policies depending on political priorities. To name but a few, having moved beyond binary classification of data according to personal or availability-based categories, existing policy and transparency requirements could be widened to create a fuller picture of processing activities, empowering users to choose between different platform models. At the same time, utilising an enhanced understanding of the value of specific data with reference to its circumstances, data collected from online platforms could be subject to more advanced data sharing rules which take into account the nature of the data and its potential uses. Bans on reuse or the requirement of clearer settings with wider applicability (such as affecting all user-generated data, rather than specific categories) could clarify user options and make easier the decision to restrict or facilitate processing. An enhanced governance and remuneration framework could empower a thriving data market economy. Obligatory third-party access could boost research capacities and make possible the closer analysis of data flows and reuse, including the precise value of different types of data. Such information, once available, promises to inform public policy, regulation, and wider discussions on how shared online spaces could and should be governed.

General Conclusion

In early autumn 2025, LinkedIn announced that data generated by users in Europe, Canada, and Hong Kong on the platform would by default be used to train and develop a number of AI models from November.²⁴³ This announcement was not the first in Europe – it followed similar statements by Meta for Instagram and Facebook earlier the same year.²⁴⁴ Nor was it a first for LinkedIn; the company had previously been forced to delay similar plans after concerns were raised by the UK Information Commissioner’s Office,²⁴⁵ and it had already introduced such data collection in other countries such as the US. The move by all of these platforms to make use of user-generated data forms part of a global race for access to data to feed the hungry appetites of LLM-based AI systems.

Platforms, whilst ranging widely in form, have become central parts of modern life. So too has the data they generate and process. This thesis has sought to examine how the processing of user-generated data from online platforms has been conceptualised and regulated under EU law, within the context of AI model development. Starting with the hypothesis that the EU regulatory framework for data was fragmented with different conceptualisations of data, impinging on the application of legal regimes, this thesis has sought to bring together an analysis of the legal framework for data and an exploration of platform practices espoused by their policies.

The systematic analysis of EU legislation concerning user-generated data in Chapter 1 revealed a heavily utilisation of conceptualisations of data based on its relation to a person and on its availability for use. Yet an economic conceptualisation was also utilised at times in the legislation. Unlike its siblings based on relation to a person and availability, however, the economic conceptualisation was under-theorised and under-present, to the detriment of many key legal regimes, most of all the DMA. There was strong fragmentation between the different conceptualisations and legal regimes, proving our hypothesis for the legal framework.

An analysis of platform policies then revealed a close alignment with the conceptualisations and legal regimes explored in the first chapter, and similarly demonstrated a lack of a developed economic conceptualisation of data. Meanwhile, within the legal regimes based on the relation of data to a person and its availability, platform policies show themselves to be struggling with the regulation of data processing in the context of AI training. Despite strong legal regimes, platform policies examined fail to describe or define bases for processing of such data under the personal data regime, whilst establishing licences with the potential for very wide use of such data under the copyright regime. Such conceptualisations and regimes, as with the legislation, were fragmented and disjointed, proving our hypothesis for platform policies.

An economic conceptualisation of data, evoked numerous times without significant development in legislation, was found to hold significant promise. The third chapter explored

²⁴³ ‘Update to Our Terms and Data Use’ (*LinkedIn*)

<<https://www.linkedin.com/help/linkedin/answer/a8059228>>.

²⁴⁴ ‘Making AI Work Harder for Europeans’ (*Meta*, 14 April 2025) <<https://about.fb.com/news/2025/04/making-ai-work-harder-for-europeans/>>.

²⁴⁵ ‘Our Statement on Changes to LinkedIn AI Data Policy’ (*Information Commissioner’s Office*, 23 September 2025) <<https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/09/our-statement-on-changes-to-linkedin-ai-data-policy/>>.

how such a conceptualisation could work both in the specific context of regulating user-generated data processing for AI training, and more generally alongside key EU data regimes. We found preliminary evidence that a strengthened economic conceptualisation of data would align well with other conceptualisations, strengthening and even simplifying the application of existing legal regimes. At the same time, it promises to provide a powerful means of regulating aspects of data collection for AI training which currently pass by parts of the regulatory framework. Such a conceptualisation would further open the door for a range of other public policies as desired.

This thesis and its methodology do have a number of limitations. Firstly, given space limits, we had to prioritise the analysis of certain legal texts and platform policies over others, with the risk of excluding relevant areas. Secondly, as a choice was made to focus on only the largest online platforms, and only a single moment in time, we cannot comment on the regulation of smaller competitors, who potentially bring innovative business models, or on the evolution of legislation and practices over time. Thirdly, both the legal and economic concepts explored here are complex and have significant pedigrees; we have often only been able to provide a brief introduction to major concepts which are still under active discussion and debate. We hope that this thesis can, however, stand as a measure of legislation and practices at this moment in time, and gesture towards what might be possible tomorrow.

At the time of writing, the European Parliament is currently considering a proposal from the Commission to amend various key texts pertaining to the regulation of data in the EU, known as the Digital Omnibus.²⁴⁶ Amongst other provisions, the proposal involves several key changes to the personal data regime of the GDPR. Namely, it proposes an increased emphasis on subjectivity in the definition of the legal category of personal data, potentially reducing the scope of the personal data category and in turn increasing that of non-personal data. It equally suggests an explicit recognition that developing an AI model can be a legitimate interest. Our research would suggest that this second proposal is already effectively the case, having identified widespread use of the legitimate interest basis in Chapter 2. Furthermore, the first proposal of a reduced scope for personal data serves to underline the value of establishing a conceptualisation of data which is not based on a patchwork of binary categorisations and legal regimes, but rather recognises a potential value in any and all data.

More widely, the Digital Omnibus stands as an opportunity to consider other modifications to the EU regulatory framework for data at an opportune moment as data collection for AI development accelerates. We propose a number of specific changes, in addition to directions for future research, detailed below.

Ultimately, we argue that a reinforced economic conceptualisation, alongside a number of other measures made possible and clarified by a consideration of the legal framework and practices using the conceptualisations of data identified, would help to contribute to the creation of a stronger regulation of data on online platforms, ready for the next novelty.

²⁴⁶ Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 2025 (Digital Omnibus)

Policy Recommendations

General Considerations:

- *For the European Parliament (EP) and European Commission (EC):* Consider the establishment of a legislative framework for data which does not use binary categories (i.e. either personal or non-personal), but instead focuses on including all data. We argue that this would help overcome a patchwork of narrow categorisations in both legislation and platform policies, bringer greater regulatory power and understandability for users.
- *For the EP, EC, national data protection authorities (NDPAs), and national competition authorities (NCAs):* Consider incorporating an economic conceptualisation of data which recognises its special properties into legislation and investigations. Such a conceptualisation could identify the complementarity and context-dependence of data, amongst other properties, and in doing so aid the identification and control of entities which acquire and process very large datasets, to the detriment of users and other developers.
- *For the EP, EC, NDPAs, and NCAs:* Consider an explicit recognition of an understanding of the value of data which is widened to include all data and variable depending on the holder, processor, and the processing involved in the legal framework and investigations. We argue that this would complement a wider economic conceptualisation of data, to ensure that all data and its effects are taken into account, important when regulating market dominance.
- *For the EP, EC, NDPAs, and NCAs:* Examine the possibility of introducing greater transparency measures (either mandatory or optional, such as in voluntary codes of conduct and recommendations) concerning how all user-generated data is processed by online providers, helping to overcome a patchwork of information which harms comprehensibility.
- *For the EC, NDPAs, and NCAs:* Explore how a more nuanced understanding of the economic value of data could influence platform business models, notably in terms of establishing a return for users without putting more vulnerable users at risk. New financing models made possible by such an understanding could equally encourage the development of new platform business models moving away from attention-based models.

Specific Measures – Economic Conceptualisation of Data:

- *For the EC and EP:* Examine the possibility of extending the prohibition on gatekeepers combining or cross-using personal data across different services (article 5(2) DMA) and using business user data across different services in competition with them (article 6(2) DMA) to include all data. This would simplify application and work more effectively in in line with the objectives of the DMA.
- *For the EC and EP:* Consider the addition of AI services to the list of CPS provided in article 2(2) DMA. This would allow for the recognition that trained systems can function across multiple different services, and could even be developed then sold, engendering potentially significant effects of concentration and power on the market.

Specific Measures – Other Conceptualisations of Data:

- *For the EC and EP:* Institute a clearer and more transparent process for the balancing of interests which are a key part of the legitimate interest legal basis in the GDPR, notably requiring publication of the balancing exercises undertaken by platforms. This would allow users and regulators to better ensure that interests are fairly balanced, even when completed by an interested party.
- *For the EC and NDPA:* Consider encouraging or even obliging platforms to provide standardised and clear information on the right to object to processing under the legitimate interest legal basis of the GDPR, including an easy capacity to refer the matter to data protection authorities. Equally consider requiring standardised clearer information on the difference between the statutory right to object and optional mitigating measures.
- *For the EC and national authorities charged with copyright enforcement:* Consider examining the extent to which platform terms are enforceable in EU law, and issue challenges to unreasonable terms. In doing so, note should be taken of the significant power imbalance between platforms with unmodifiable “boilerplate” contracts and users, who may feel obliged to agree in order to access platforms which form important parts of modern life.
- *For the EC and EP:* Examine how the development of so-called “*People’s Terms of Service*”, allowing individuals to debate and construct the platform terms they prefer, can be encouraged, notably regarding their acceptance by platforms. Consider instituting a right to choose between multiple terms of service by addition to the DSA.

Recommended Directions for Future Research:

- *For researchers:* Specific research into the economic tendencies affecting the collection and processing of large amounts of data in the specific context of online platforms and AI development would allow for better-informed specific legal categories to be established using an economic conceptualisation of data. This would allow for the identification of factors affecting the value of data in different contexts and circumstances, feeding into discussions on governance, returns, and fairness for users. More widely, this can assist in ensuring greater competition on the market, and encourage alternative models for digital services to the advantage of all.
- *For researchers:* Research into the effectiveness of different forms and methods of pseudonymisation, anonymisation, and other data protection operations within the specific context of LLMs and AI models would allow for a more informed regulation to emerge within the personal and copyright data regimes, aimed at ensuring privacy and ideas are protected.
- *For researchers:* An examination of the mechanisms by which platforms can be encouraged to institute alternative terms of use, including potentially a “*People’s Terms of Service*”, would help to reduce the asymmetry of power between the two parties and aid users in having more choice over the functioning of platforms and the processing of their data.

Bibliography

Legislation

Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') 2000 (OJ L 178) 1

Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data 2000 (OJ L 008)

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society 2001 (OJ L 167) 10

Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) 2002 (OJ L 201) 37

Charter of Fundamental Rights of the European Union 2012 (OJ C 326) 391

Consolidated version of the Treaty on the Functioning of the European Union 2012 (OJ C 326) 47

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) 2016 (OJ L 119 452016)

Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure (Text with EEA relevance) 2016 (OJ L 157) 1

Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA 2017 (OJ L 88) 6

Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union (Text with EEA relevance.) 2018 (OJ L 303) 59

Directive (EU) 2018/1972 of the European Parliament and of the Council of 11 December 2018 establishing the European Electronic Communications Code (Recast) Text with EEA relevance. 2018 (OJ L 321) 36

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.) 2019 (OJ L 13) 92

Directive (EU) 2019/770 of the European Parliament and of the Council of 20 May 2019 on certain aspects concerning contracts for the supply of digital content and digital services Text with EEA relevance. 2019 (OJ L 136) 1

Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast) 2019 (OJ L 172) 56

Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services (Text with EEA relevance) 2019 (OJ L 186) 57

European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies (2020/2015(INI)) 2020 (OJ C 404) 129

Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online (Text with EEA relevance) 2021 (OJ L 172)

Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) (Text with EEA relevance) 2022 (OJ L 152)

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance) 2022 (OJ L 265)

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) 2022 (OJ L 277) 1

Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) (Text with EEA relevance) 2023 (OJ L 2023/2854)

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) 2024 (OJ L 2024/1689)

Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR)

Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus) 2025

Case Law

Esch-Leonhardt and Others v ECB (ECJ) [2004] ECJ Case T-320/02

Patrick Breyer v Bundesrepublik Deutschland [2016] ECJ Case C-582/14

Peter Nowak v Data Protection Commissioner [2017] ECJ Case C-434/16

Bundeskartellamt, Facebook [2019] Case B6-22/16, Decision of 6 February 2019.

https://www.bundeskartellamt.de/SharedDocs/Entscheidung/EN/Fallberichte/Missbrauchsaufsicht/2019/B6-22-16.pdf?__blob=publicationFile&v=3

Mircom International Content Management & Consulting (MICM) Limited v Telenet BVBA [2021] ECJ Case C-597/19

Meta Platforms Inc and Others v Bundeskartellamt [2023] ECJ Case C-252/21

Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens [2024] ECJ Case C-621/22

European Data Protection Supervisor v Single Resolution Board [2025] ECJ Case C-413/23
P

Platform Policies

‘About End-to-End Encryption’ (*WhatsApp*) <<https://faq.whatsapp.com/820124435853543>> accessed 15 March 2026

‘Additional Information about Data Processing’ (*X*) <<https://help.x.com/en/rules-and-policies/data-processing-legal-bases>> accessed 15 March 2026

‘AI at Pinterest’ (*Pinterest*) <<https://help.pinterest.com/en-gb/article/ai-at-pinterest>> accessed 15 March 2026

‘AI Features Privacy FAQ’ (*LinkedIn*, 2025)
<<https://www.linkedin.com/help/linkedin/answer/a5538339>> accessed 15 March 2026

‘App Store & Privacy’ (*Apple*, 12 December 2025)
<<https://www.apple.com/legal/privacy/data/en/app-store/>> accessed 12 March 2026

‘Apple ID & Privacy’ (*Apple*, 23 July 2025) <<https://support.apple.com/en-us/120320>> accessed 12 March 2026

‘Apple Media Services Terms and Conditions’ (*Apple*, 15 September 2025)
<<https://www.apple.com/uk/legal/internet-services/itunes/uk/terms.html>> accessed 12 March 2026

‘Applebot Model Training and Individual Privacy Rights’ (*Apple*, 23 July 2025)
<<https://support.apple.com/en-us/120320>> accessed 12 March 2026

‘Children’s Features Privacy Notice’ (*Google*, 31 May 2022)
<https://assistant.google.com/privacy-notice-childrens-features/?hl=en_GB> accessed 13 March 2026

‘Conditions of Use’ (*Amazon*, 28 November 2025)
 <<https://www.amazon.co.uk/gp/help/customer/display.html?nodeId=GLSBYFE9MGKKQX>
[XM](#)> accessed 12 March 2026

‘Cookies Notice’ (*Amazon*, 20 February 2025)
 <<https://www.amazon.co.uk/gp/help/customer/display.html?nodeId=G64JFZVFDY66XG9K>
 > accessed 12 March 2026

‘Data Processing Legal Bases’ (*X*) <<https://help.x.com/en/rules-and-policies/data-processing-legal-bases>> accessed 15 March 2026

‘EEA/UK Privacy Notice’ (*Snapchat*, 6 November 2023)
 <<https://values.snap.com/privacy/privacy-policy/eea-uk-privacy-notice>> accessed 15 March 2026

‘EU AI Terms’ (*Meta*, 5 January 2026) <<https://www.facebook.com/legal/eu-ai-terms>>
 accessed 15 March 2026

‘Free Membership Agreement’ (*AliExpress*, 1 March 2025) <https://terms.alicdn.com/legal-agreement/terms/suit_bu1_alieexpress/suit_bu1_alieexpress202204182115_37406.html>
 accessed 12 March 2026

‘Gemini Apps Privacy Hub’ (*Google*, 10 March 2026)
 <<https://support.google.com/gemini/answer/13594961?hl=en>> accessed 15 March 2026

‘Gemini Terms of Service’ (*Google*, 10 March 2026)
 <<https://support.google.com/gemini/answer/13594961?hl=en>> accessed 13 March 2026

‘Generative AI on Snapchat’ (*Snapchat*) <<https://help.snapchat.com/hc/en-us/articles/25494876770580-Generative-AI-on-Snapchat>> accessed 15 March 2026

‘Generative AI Privacy Information’ (*Meta*) <<https://www.facebook.com/privacy/genai>>
 accessed 15 March 2026

‘Generative AI Use Policy’ (*Google*, 17 December 2024)
 <<https://policies.google.com/terms/generative-ai/use-policy?hl=en-GB&fg=1>> accessed 13 March 2026

‘Google Drive Terms of Service’ (*Google*, 22 October 2025)
 <<https://www.google.com/drive/terms-of-service/>> accessed 13 March 2026

‘Google Maps Terms of Service’ (*Google*, 4 June 2025)
 <https://www.google.com/help/terms_maps/?hl=en_GB> accessed 13 March 2026

‘Google Play Terms of Service’ (*Google*, 6 October 2025)
 <<https://play.google.com/about/play-terms/index.html>> accessed 13 March 2026

‘How ChatGPT and Our Foundation Models Are Developed’ (*OpenAI*, 30 April 2025)
 <<https://openai.com/en-GB/policies/how-chatgpt-and-our-foundation-models-are-developed/>> accessed 29 March 2026

‘Instagram Privacy Policy’ (*Meta*, 16 December 2025)
 <<https://privacycenter.instagram.com/policy/>> accessed 15 March 2026

‘Instagram Terms of Use’ (*Meta*, 1 January 2025) <<https://help.instagram.com/termsfuse>> accessed 15 March 2026

‘Internet Services Terms and Conditions’ (*Apple*, 20 November 2009) <<https://www.apple.com/uk/legal/internet-services/terms/site.html>> accessed 12 March 2026

‘Legal Notice’ (*Zalando*) <<https://www.zalando.co.uk/legal-notice/>> accessed 21 March 2026

‘LinkedIn European Regional Privacy Notice’ (*LinkedIn*, 3 November 2025) <<https://www.linkedin.com/legal/privacy/eu>> accessed 15 March 2026

‘Making AI Work Harder for Europeans’ (*Meta*, 14 April 2025) <<https://about.fb.com/news/2025/04/making-ai-work-harder-for-europeans/>> accessed 18 March 2026

‘Microsoft DMA Compliance Report 2026: Annex 11 – LinkedIn (Online Social Networking Service)’ (2026) <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/mscle/documents/presentations/Microsoft_DMA_Compliance_Report%e2%80%9393LinkedIn_Section_2_Annex.pdf>

‘Perplexity on Snapchat’ (*Snapchat*) <<https://help.snapchat.com/hc/en-us/articles/43680949549844-What-is-Perplexity-on-Snapchat-and-how-do-I-use-it>> accessed 15 March 2026

‘Privacy and Cookie Policy’ (*Temu*, 2024) <<https://www temu.com/privacy-and-cookie-policy.html>> accessed 21 March 2026

‘Privacy and Cookie Policy France’ (*Temu*, 11 October 2025) <<https://www temu.com/fr-en/privacy-and-cookie-policy.html>> accessed 21 March 2026

‘Privacy by Product’ (*Snapchat*) <<https://values.snap.com/privacy/privacy-by-product>> accessed 15 March 2026

‘Privacy FAQ for Microsoft Copilot’ (*Microsoft*) <<https://support.microsoft.com/en-gb/topic/privacy-faq-for-microsoft-copilot-27b3a435-8dc9-4b55-9a4b-58eeb9647a7f#ID0EFF>> accessed 15 March 2026

‘Privacy Notice’ (*XVideos*, 8 August 2022) <<https://info.xvideos.net/legal/privacynotice>> accessed 21 March 2026

‘Privacy Notice’ (*Amazon*, 1 August 2024) <https://www.amazon.co.uk/gp/help/customer/display.html?nodeId=GX7NJQ4ZB8MHFRNJ&ref=footer_privacy> accessed 12 March 2026

‘Privacy Notice’ (*Shein*, 18 November 2025) <<https://www.shein.co.uk/Consumer-Privacy-Notice-s-101.html>> accessed 15 March 2026

‘Privacy Policy’ (*XNXX*, 1 September 2022) <<https://info xnxx.com/legal/privacy>> accessed 21 March 2026

‘Privacy Notice’ (*XVideos*, 1 June 2023) <<https://info.xvideos.net/legal/privacy>> accessed 21 March 2026

‘Privacy Notice’ (*Snapchat*, 7 April 2025) <<https://values.snap.com/privacy/privacy-policy>> accessed 15 March 2026

‘Privacy Notice’ (*Pinterest*, 30 April 2025) <<https://policy.pinterest.com/en/privacy-policy>> accessed 15 March 2026

‘Privacy Notice’ (*Apple*, 30 July 2025) <<https://www.apple.com/legal/privacy/en-ww/>> accessed 12 March 2026

‘Privacy Notice’ (*Wikimedia Foundation*, 3 September 2025) <https://foundation.wikimedia.org/wiki/Policy:Privacy_policy> accessed 21 March 2026

‘Privacy Notice’ (*AliExpress*, 14 October 2025) <https://terms.alicdn.com/legal-agreement/terms/suit_bu1_alieexpress/suit_bu1_alieexpress201909171350_82407.html> accessed 12 March 2026

‘Privacy Notice’ (*LinkedIn*, 3 November 2025) <<https://www.linkedin.com/legal/privacy-policy>> accessed 15 March 2026

‘Privacy Notice’ (*Google*, 11 December 2025) <<https://policies.google.com/privacy?hl=en-GB&fg=1>> accessed 13 March 2026

‘Privacy Notice’ (*X*, 15 January 2026) <<https://x.com/en/privacy>> accessed 15 March 2026

‘Privacy Notice’ (*Pornhub*, 30 January 2026) <<https://www.pornhub.com/information/privacy>> accessed 12 March 2026

‘Privacy Notice’ (*Meta*, 4 March 2026) <<https://www.facebook.com/privacy/policy/>> accessed 15 March 2026

‘Privacy Notice’ (*Zalando*, 9/25) <<https://www.zalando.co.uk/zalando-privacy-policy/>> accessed 21 March 2026

‘Privacy Policy (EEA)’ (*TikTok*, 30 November 2025) <<https://www.tiktok.com/legal/page/eea/privacy-policy/en>> accessed 15 March 2026

‘Privacy Statement’ (*Booking.com*, December 2025) <<https://www.booking.com/content/privacy.en-gb.html>> accessed 12 March 2026

‘Privacy Statement’ (*Microsoft*, 03/26) <<https://www.microsoft.com/en-gb/privacy/privacystatement>> accessed 15 March 2026

‘Search Generative Experience Terms’ (*Google*) <<https://support.google.com/websearch/answer/14108842?hl=en-GB#zippy=>> accessed 13 March 2026

‘Search Terms of Service’ (*Google*, 15 November 2023) <https://www.google.com/search/content/terms/?hl=en_GB> accessed 13 March 2026

‘Services Agreement’ (*Microsoft*, 30 July 2025) <<https://www.microsoft.com/en-gb/servicesagreement/>> accessed 16 March 2026

‘Terms and Conditions’ (*Booking.com*, 15 September 2025) <<https://www.booking.com/content/terms.en-gb.html>> accessed 12 March 2026

‘Terms and Conditions’ (*Shein*, 11 February 2026) <<https://www.shein.co.uk/Terms-and-Conditions-a-399.html>> accessed 15 March 2026

‘Terms and Conditions for Platform Transactions by EU/EEA/UK Consumers’ (*AliExpress*, 1 August 2025) <https://cdn.contract.alibaba.com/terms/EU_EE_UK_platform_service_agreement/20250514112355975/20250514112355975.html> accessed 12 March 2026

‘Terms of Service’ (*XNXX*, 17 February 2024) <<https://info.xnxx.com/legal/tos>> accessed 21 March 2026

‘Terms of Service’ (*XVideos*, 17 February 2024) <<https://info.xvideos.net/legal/tos>> accessed 21 March 2026

‘Terms of Service’ (*Google*, 22 May 2024) <<https://policies.google.com/terms?hl=en-GB&fg=1>> accessed 13 March 2026

‘Terms of Service’ (*Snapchat*, 7 April 2025) <<https://www.snap.com/terms>> accessed 16 March 2026

‘Terms of Service’ (*Pinterest*, 30 April 2025) <<https://policy.pinterest.com/en-gb/terms-of-service>> accessed 15 March 2026

‘Terms of Service’ (*Pornhub*, 19 November 2025) <<https://www.pornhub.com/information/terms>> accessed 12 March 2026

‘Terms of Service’ (*X*, 15 January 2026) <<https://x.com/en/tos>> accessed 15 March 2026

‘Terms of Service’ (*Meta*, 4 March 2026) <<https://www.facebook.com/terms/?lang=en>> accessed 15 March 2026

‘Terms of Service (EEA)’ (*WhatsApp*, 16 February 2024) <<https://www.whatsapp.com/legal/terms-of-service-eea>> accessed 15 March 2026

‘Terms of Service (EEA)’ (*TikTok*, August 2025) <<https://www.tiktok.com/legal/page/eea/terms-of-service/en>> accessed 15 March 2026

‘Terms of Use’ (*Wikimedia Foundation*, 7 June 2023) <https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use> accessed 21 March 2026

‘Terms of Use’ (*AliExpress*, 1 March 2025) <https://terms.alicdn.com/legal-agreement/terms/suit_bu1_alieexpress/suit_bu1_alieexpress202204182115_66077.html> accessed 12 March 2026

‘Terms of Use (UK)’ (*Temu*, 11 October 2025) <<https://www.temu.com/uk/terms-of-use.html>> accessed 21 March 2026

‘UK AI Terms’ (*Meta*, 18 December 2025) <<https://www.facebook.com/legal/uk-ai-terms>> accessed 15 March 2026

‘Update to Our Terms and Data Use’ (*LinkedIn*) <<https://www.linkedin.com/help/linkedin/answer/a8059228>> accessed 18 March 2026

‘User Agreement’ (*LinkedIn*, 3 November 2025) <<https://www.linkedin.com/legal/user-agreement/>> accessed 16 March 2026

‘User Contributed Content Policy’ (*Google*)
<<https://support.google.com/contributionpolicy/answer/7422880?hl=en-GB>> accessed 13 March 2026

‘WhatsApp DMA Notice for Non-Users’ (*WhatsApp*, 1 October 2025)
<<https://www.whatsapp.com/legal/dma-notice-non-users>> accessed 15 March 2026

‘WhatsApp Privacy Policy (EEA)’ (*WhatsApp*, 16 February 2024)
<<https://www.whatsapp.com/legal/privacy-policy-eea>> accessed 15 March 2026

‘WhatsApp Privacy Policy (UK)’ (*WhatsApp*, 20 March 2025)
<<https://www.whatsapp.com/legal/privacy-policy-uk>> accessed 15 March 2026

‘WhatsApp Terms of Service (EEA)’ (*WhatsApp*, 16 February 2024)
<<https://www.whatsapp.com/legal/terms-of-service-eea>> accessed 15 March 2026

‘WhatsApp Terms of Service (UK)’ (*WhatsApp*, 17 March 2025)
<<https://www.whatsapp.com/legal/terms-of-service-uk>> accessed 15 March 2026

‘Workspace for Education Privacy Notice’ (*Google*)
<https://workspace.google.com/terms/education_privacy/> accessed 13 March 2026

‘YouTube Terms of Service’ (*Google*, 9 January 2026)
<https://www.youtube.com/t/terms?hl=en&override_hl=1> accessed 13 March 2026

Secondary Sources

‘Annual Scholarly Publications on Artificial Intelligence’ (*Our World in Data*)
<<https://ourworldindata.org/grapher/annual-scholarly-publications-on-artificial-intelligence>>
accessed 10 April 2026

‘data’ (Cambridge Dictionary) <<https://dictionary.cambridge.org/dictionary/english/data>>
accessed 14 April 2026

‘Generative AI and the EUDPR: Orientations for Ensuring Data Protection Compliance When Using Generative AI Systems’ (European Data Protection Supervisor 2025)
<https://www.edps.europa.eu/system/files/2025-10/25-10_28_revised_genai_orientations_en.pdf>

‘Generative Artificial Intelligence: The Autorité Issues Its Opinion on the Competitive Functioning of the Sector’ (*Autorité de la concurrence*, 28 June 2024)
<<https://www.autoritedelaconcurrence.fr/en/press-release/generative-artificial-intelligence-autorite-issues-its-opinion-competitive>> accessed 11 April 2026

‘Guidelines 01/2025 on Pseudonymisation’ (European Data Protection Board 2025)
<https://www.edpb.europa.eu/system/files/2025-01/edpb_guidelines_202501_pseudonymisation_en.pdf>

‘Guidelines 1/2024 on Processing of Personal Data Based on Article 6(1)(f) GDPR’ (European Data Protection Board 2024) <https://www.edpb.europa.eu/system/files/2024-10/edpb_guidelines_202401_legitimateinterest_en.pdf>

‘Opinion 08/2024 on Valid Consent in the Context of Consent or Pay Models Implemented by Large Online Platforms’ (European Data Protection Board 2024) <https://www.edpb.europa.eu/system/files/2024-04/edpb_opinion_202408_consentorpay_en.pdf>

‘Introduction to Anonymisation’ (*Information Commissioner’s Office*) <<https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/introduction-to-anonymisation/>> accessed 13 March 2026

‘L’intérêt légitime : comment fonder un traitement sur cette base légale ?’ (CNIL, 2 December 2019) <<https://www.cnil.fr/fr/les-bases-legales/interet-legitime>> accessed 17 April 2026

‘Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models’ (European Data Protection Board 2024) <https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf>

‘Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models’ (European Data Protection Board 2024) <https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf>

‘Our Statement on Changes to LinkedIn AI Data Policy’ (*Information Commissioner’s Office*, 23 September 2025) <<https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/09/our-statement-on-changes-to-linkedin-ai-data-policy/>> accessed 18 March 2026

‘Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World’ (Stanford University Human-Centered Artificial Intelligence 2024) <<https://hai.stanford.edu/policy/white-paper-rethinking-privacy-ai-era-policy-provocations-data-centric-world>>

‘Supervision of the Designated Very Large Online Platforms and Search Engines under DSA’ (European Commission, 1 April 2026) <<https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses>> accessed 10 April 2026

‘What Are the Conditions for Processing?’ (*Information Commissioner’s Office*, 18 December 2023) <<https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-are-the-conditions-for-processing/>> accessed 22 March 2026

‘Wikipedia Kaggle Dataset Using Structured Contents Snapshot’ (*Wikimedia Enterprise*, 16 April 2025) <<https://enterprise.wikimedia.com/blog/kaggle-dataset/>> accessed 22 March 2026

Adam D, ‘The Intelligence Illusion: Why AI Isn’t as Smart as It Is Made out to Be’ (2026) 651 *Nature* 877 <<https://doi.org/10.1038/d41586-026-00882-5>>

Ahmed A and others, ‘Extracting Books from Production Language Models’ (arXiv, 6 January 2026) <<https://doi.org/10.48550/arXiv.2601.02671>>

Benoliel U and Becher S, ‘The Duty to Read the Unreadable’ (Social Science Research Network, 11 January 2019) <<https://doi.org/10.2139/ssrn.3313837>>

Bostoen F and Krämer J, ‘How Future-Proof Is the DMA? A Case Study Of AI Agents’ [2026] *Journal of Competition Law & Economics* 1 <<https://doi.org/10.1093/joclec/nhag005>>

Burden J and others, ‘Legal Aspects of Access to Human-Generated Data and Other Essential Inputs for AI Training’ (Social Science Research Network, 2 December 2024) <<https://doi.org/10.2139/ssrn.5045155>>

Carlini N and others, ‘Extracting Training Data from Large Language Models’ [2021] 30th USENIX Security Symposium 2633

Carlini N and others, ‘The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks’ *Proceedings of the 28th USENIX Conference on Security Symposium* (USENIX Association 2019)

Carovano G and Finck M, ‘Regulating Data Intermediaries: The Impact of the Data Governance Act on the EU’s Data Economy’ (2023) 50 *Computer Law & Security Review* 105830 <<https://doi.org/10.1016/j.clsr.2023.105830>>

Christophe B, Giron A and Verin G, ‘A Comparative Analysis with Machine Learning of Public Data Governance and AI Policies in the European Union, United States, and China’ (2023) 13 *Journal of Intelligence Studies in Business* 61 <<https://doi.org/10.37380/jisib.v13i2.1084>>

Coyle D and Manley A, ‘What Is the Value of Data? A Review of Empirical Methods’ (Bennett Institute for Public Policy 2022) <https://www.bennettschool.cam.ac.uk/wp-content/uploads/2022/07/policy-brief_what-is-the-value-of-data.pdf>

Coyle S, ‘Hart, Raz and the Concept of a Legal System’ (2002) 21 *Law and Philosophy* 275

Delacroix S and Lawrence ND, ‘Bottom-up Data Trusts: Disturbing the “One Size Fits All” Approach to Data Governance’ (2019) 9 *International Data Privacy Law* 236 <<https://doi.org/10.1093/idpl/ipz014>>

Desmarais A, ‘German Non-Profits Win Data Access Court Case against X’ *euronews* (5 February 2025) <<https://www.euronews.com/next/2025/02/07/we-must-defend-ourselves-german-non-profits-sue-x-for-access-to-public-data>> accessed 17 February 2026

Dove ES and Chen J, ‘What Does It Mean for a Data Subject to Make Their Personal Data “Manifestly Public”? An Analysis of GDPR Article 9(2)(e)’ (2021) 11 *International Data Privacy Law* 107 <<https://doi.org/10.1093/idpl/ipab005>>

Ducato R and Strowel A, ‘Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out’ (2021) 43 *European Intellectual Property Review* 322

European Data Protection Supervisor, ‘Guidance for Risk Management of Artificial Intelligence Systems’ (2025) <https://www.edps.europa.eu/data-protection/our-work/publications/guidelines/2025-11-11-guidance-risk-management-artificial-intelligence-systems_en>

Finck M and Pallas F, ‘They Who Must Not Be Identified—Distinguishing Personal from Non-Personal Data under the GDPR’ (2020) 10 International Data Privacy Law <<https://academic.oup.com/idpl/article/10/1/11/5802594>>

Fletcher A and others, ‘The Effective Use of Economics in the EU Digital Markets Act’ (2024) 20 Journal of Competition Law & Economics 1 <<https://doi.org/10.1093/joclec/nhad018>>

Gracz K and de Filippi P, ‘Regulatory Failure of Copyright Law through the Lenses of Autopoietic Systems Theory’ [2014] International Journal of Law and Information Technology 46

Hacker P, ‘A Legal Framework for AI Training Data—from First Principles to the Artificial Intelligence Act’ (2021) 13 Law, Innovation and Technology 257 <<https://doi.org/10.1080/17579961.2021.1977219>>

Hacker P, Cordes J and Rochon J, ‘Regulating Gatekeeper Artificial Intelligence and Data: Transparency, Access and Fairness under the Digital Markets Act, the General Data Protection Regulation and Beyond’ (2024) 15 European Journal of Risk Regulation 49 <<https://doi.org/10.1017/err.2023.81>>

Hartzog W, ‘Fighting Facebook, a Campaign for a People’s Terms of Service’ (*Stanford Center for Internet and Society*, 22 May 2013) <<https://cyberlaw.stanford.edu/blog/2013/05/fighting-facebook-campaign-peoples-terms-service/>> accessed 31 March 2026

Häuselmann A and others, ‘EU Law and Emotion Data’ (IEEE 2023) <<https://doi.org/10.1109/ACII59096.2023.10388181>>

Helberger N, Borgesius FZ and Reyna A, ‘The Perfect Match? A Closer Look at the Relationship Between EU Consumer Law and Data Protection Law’ (2017) 54 Common Market Law Review 1427

ICC, ‘ICC Policy Primer on Non-Personal Data’ (International Chamber of Commerce 2023) <iccwbo.org/news-publications/policies-reports/ICC-policy-primer-on-non-personal-data/>

Ilari A, ‘Le cas TADATA’ (*DL4T - Université Côte d’Azur*) <<https://droit.univ-cotedazur.fr/dl4t/le-cas-tadata>> accessed 10 April 2026

Kerber W, ‘Rights on Data: The EU Communication “Building a European Data Economy” from an Economic Perspective’ (University of Marburg 2017) MAGKS Joint Discussion Paper Series in Economics 35–2017 <<https://www.econstor.eu/handle/10419/174331>>

Kim H, ‘How Much of the Market Is AI?’ *Financial Times* (18 November 2025) <<https://www.ft.com/content/1edc4113-4282-474d-928b-4b84b7f2044e?syn-25a6b1a6=1>> accessed 10 April 2026

- King J and others, ‘User Privacy and Large Language Models: An Analysis of Frontier Developers’ Privacy Policies’ (2025) 8 Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 1465 <<https://doi.org/10.1609/aies.v8i2.36646>>
- Kirkwood M, ‘Will the EU Designate AI Under the Digital Markets Act?’ (*Tech Policy Press*, 26 September 2025) <<https://www.techpolicy.press/will-the-eu-designate-ai-under-the-digital-markets-act/>> accessed 9 April 2026
- Liu Y and others, ‘Evaluating LLM-Based Personal Information Extraction and Countermeasures’ [2025] 34th USENIX Security Symposium <<https://www.usenix.org/system/files/usenixsecurity25-liu-yupei.pdf>>
- Malgieri G and Custers B, ‘Pricing Privacy – the Right to Know the Value of Your Personal Data’ (2018) 34 Computer Law & Security Review 289 <<https://doi.org/10.1016/j.clsr.2017.08.006>>
- Meyn J and others, ‘Consequences of Platforms’ Remuneration Models for Digital Content: Initial Evidence and a Research Agenda for Streaming Services’ (2023) 51 Journal of the Academy of Marketing Science 114 <<https://doi.org/10.1007/s11747-022-00875-6>>
- Mills S, ‘Who Owns the Future? Data Trusts, Data Commons, and the Future of Data Ownership’ (Social Science Research Network, 24 September 2019) <<https://doi.org/10.2139/ssrn.3437936>>
- Mitchell J, Ker D and Leshner M, ‘Measuring the Economic Value of Data’ (OECD Publishing 2021) <<https://doi.org/10.1787/f46b3691-en>>
- Møller AG and others, ‘The Impact of Generative AI on Social Media: An Experimental Study’ (2026) 16 Nature Scientific Reports 9376 <<https://doi.org/10.1038/s41598-026-40110-8>>
- Nasr M and others, ‘Scalable Extraction of Training Data from (Production) Language Models’ (November 2023) <<https://doi.org/10.48550/arXiv.2311.17035>>
- Neally D and Leong B, ‘Copyrights and Privacy: What Is the Irrevocable License and Is It Really a Privacy Concern?’ (*Future of Privacy Forum*, 30 July 2019) <<https://fpf.org/blog/copyrights-and-privacy-what-is-the-irrevocable-license-and-is-it-really-a-privacy-concern/>> accessed 22 March 2026
- Nielsen SPP, ‘Introduction: A Special Issue on Niklas Luhmann’s Systems Theory and Law’ (2024) 14 Oñati Socio-Legal Series 1206 <<https://doi.org/10.35295/osls.iisl.1855>>
- Nobles R and Schiff D, ‘Using Systems Theory to Study Legal Pluralism: What Could Be Gained?’ (2012) 46 Law & Society Review 265
- Nobles R and Schiff D, *Observing Law Through Systems Theory* (1st edn, Bloomsbury Publishing Plc 2012).
- Novelli C and others, ‘Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity’ (2024) 55 Computer Law & Security Review 106066 <<https://doi.org/10.1016/j.clsr.2024.106066>>

Novović M, ‘Licensing of User-Generated Content: How Online Platforms Exploit Users’ Copyrighted Works’ (University of Oslo 2017) <<https://milos.no/wp-content/uploads/2023/07/Milos-Novovic-PhD-Thesis-1.pdf>>

Novović M, ‘User-Generated Content: How Broad Licensing Terms Threaten the Web’ *The Responsibilities of Online Service Providers* (Springer International Publishing 2017) <https://doi.org/10.1007/978-3-319-47852-4_11>

NOYB, ‘Digital Omnibus: First Analysis of Select GDPR and ePrivacy Proposals by the Commission’ (2026) <<https://noyb.eu/en/digital-omnibus-report-v2-analysis-select-gdpr-and-eprivacy-proposals-commission>>

NOYB, ‘Pay or Okay: How Companies Make You Pay for Privacy’ (2025) <<https://noyb.eu/en/noybs-pay-or-okay-report-how-companies-make-you-pay-privacy>> accessed 17 February 2026

OECD, ‘An Introduction to Online Platforms and Their Role in the Digital Transformation’ (OECD Publishing 2019) <<https://doi.org/10.1787/53e5f593-en>>

OECD, ‘Artificial Intelligence, Data and Competition’ [2024] OECD Artificial Intelligence Papers <<https://doi.org/10.1787/e7e88884-en>>

OECD, ‘Mapping Relevant Data Collection Mechanisms for AI Training’ (OECD Publishing 2025) OECD Artificial Intelligence Papers <<https://doi.org/10.1787/3264cd4c-en>>

OECD, ‘OECD Framework for the Classification of AI Systems’, vol 323 (2022) OECD Digital Economy Papers 323 <<https://doi.org/10.1787/cb6d9eca-en>>

Purtova N and van Maanen G, ‘Data as an Economic Good, Data as a Commons, and Data Governance’ (2024) 16 Law, Innovation and Technology 1 <<https://doi.org/10.1080/17579961.2023.2265270>>

Purtova N, ‘The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law’ (2018) 10 Law, Innovation and Technology 40 <<https://doi.org/10.1080/17579961.2018.1452176>>

Quintais JP, De Gregorio G and Magalhães JC, ‘How Platforms Govern Users’ Copyright-Protected Content: Exploring the Power of Private Ordering and Its Implications’ (2023) 48 Computer Law & Security Review 105792 <<https://doi.org/10.1016/j.clsr.2023.105792>>

Riis N, ‘Shaping the Field of EU Data Law’ (2023) 14 Journal of Intellectual Property, Information Technology and Electronic Commerce Law <<https://doi.org/10.2139/ssrn.4533000>>

Ruscheimer H, ‘Generative AI and Data Protection’ (2025) 1 Cambridge Forum on AI: Law and Governance e6 <<https://doi.org/10.1017/cfl.2024.2>>

Ruscheimer H, ‘Squaring the Circle’ (*Verfassungsblog*, 7 April 2023) <<https://doi.org/10.17176/20230407-190249-0>> accessed 10 April 2026

Sag M, ‘Copyright Safety for Generative AI’ (2023) 61 Houston Law Review 295

Schweitzer H and Metzger A, 'Data Access under the Draft Data Act, Competition Law and the DMA: Opening the Data Treasures for Competition and Innovation?' (2023) 72 GRUR International 337 <<https://doi.org/10.1093/grurint/ikad012>>

Shakespeare, William, *Hamlet* (1601) [2020], Folger Shakespeare Library.

Shumailov I and others, 'AI Models Collapse When Trained on Recursively Generated Data' (2024) 631 Nature 755 <<https://doi.org/10.1038/s41586-024-07566-y>>

Sophia DM, 'Wikipedia Owner Signs on Microsoft, Meta in AI Content Training Deals' *Reuters* (15 January 2026) <<https://www.reuters.com/business/retail-consumer/wikipedia-owner-signs-microsoft-meta-ai-content-training-deals-2026-01-15/>> accessed 22 March 2026

Stalla-Bourdillon S, 'Identifiability, as a Data Risk: Is a Uniform Approach to Anonymisation About to Emerge in the EU?' (2025) 16 European Journal of Risk Regulation 1456 <<https://doi.org/10.1017/err.2025.16>>

Stalla-Bourdillon S, 'The EDPB 01/2025 Guidelines on Pseudonymisation: A Step in the Right Direction?' [2025] European Law Blog <<https://doi.org/10.21428/9885764c.3d9978a3>> accessed 5 April 2026

Thomas D and Gross A, 'UK to Delay Difficult Decisions on AI Copyright Rules' *Financial Times* (6 March 2026) <<https://www.ft.com/content/e759a712-eddf-4bdd-b4d9-03446f8c6545?syn-25a6b1a6=1>> accessed 22 March 2026

Thomas D, 'UK Creative Sector Seeks Copyright Protection from AI Companies' *Financial Times* (16 December 2024) <<https://www.ft.com/content/c697f9be-57ac-43e4-a351-4b096e2136d0?syn-25a6b1a6=1>> accessed 13 April 2026

Toptchiyska D, 'Navigating The GDPR-DSA Nexus: Regulating Personal Data In Social Media and Search Engines' (Social Science Research Network, 25 May 2024) <<https://doi.org/10.2139/ssrn.4948546>>

Wachter S, 'Data Protection in the Age of Big Data' (2019) 2 Nature Electronics 6 <<https://doi.org/10.1038/s41928-018-0193-y>>

Wauters E, Lievens E and Valcke P, 'Towards a Better Protection of Social Media Users: A Legal Perspective on the Terms of Use of Social Networking Sites' (2014) 22 International Journal of Law and Information Technology 254

Willcocks L and Currie W, 'AI Is a Misleading Tale of Hope and Fear' (*LSE Business Review*, 20 March 2026) <<https://blogs.lse.ac.uk/businessreview/2026/03/20/ai-is-a-misleading-tale-of-hope-and-fear/>> accessed 14 April 2026

Wolf M, 'Playing "Whack-a-Mole" with Meta over My Fraudulent Avatars' *Financial Times* (25 April 2025) <<https://www.ft.com/content/80d9c4f8-653d-4f8c-89b8-b4d1e183a134?syn-25a6b1a6=1>> accessed 6 April 2026

Yasar AG and others, 'AI and the EU Digital Markets Act: Addressing the Risks of Bigness in Generative AI' (arXiv, 7 July 2023) <<https://doi.org/10.48550/arXiv.2308.02033>>

Zhan X, de València P and Seymour W, 'Malicious LLM-Based Conversational AI Makes Users Reveal Personal Information'

Annex 1: DSA-designated VLOPs and VLOSEs and their Policies

<u>Provider</u>	<u>Designated Service(s)</u>	<u>Description</u>	<u>Avg. monthly users (millions)</u>	<u>Policies examined</u>
AliExpress International (Netherlands) B.V.	AliExpress	Retail platform	104.3	Terms and Conditions for Platform Transactions Privacy Policy Free Membership Agreement Terms of Use
Amazon EU S.à.r.l.	Amazon Store	Retail platform	181.3	Conditions of Use Cookies Notice Privacy Notice Interest-Based Ads
Apple Distribution International Limited	App Store	App store	123	Privacy Policy Apple ID & Privacy App Store & Privacy Apple Media Terms and Conditions Internet Services Terms and Conditions Applebot model training and individual privacy rights
Aylo Freesites Ltd.	Pornhub	Pornography video-sharing	Above 45	Privacy Policy Terms of Service
Booking.com B.V.	Booking.com	Accommodation marketplace	Above 45	Privacy Statement Terms and Conditions
Google Ireland Ltd.	Google Search	Search engine	364	Privacy Policy Children's Features Privacy Notice Workspace for Education Privacy Notice Terms of Service Generative AI Use Policy Google Play Terms of Service Search Generative Experience Terms Search Terms of Service Google Drive Terms of Service
	Google Play	App store	284.6	
	Google Maps	Map platform	275.6	
	Google Shopping	Shopping platform	70.8	
	YouTube	Video-sharing platform	416.6	

				Google Maps Terms of Service User Contributed Content Policy YouTube Terms of Service Gemini Terms of Service
Infinite Styles Services Co, Ltd	Shein	Retail platform	108	Terms and Conditions Privacy Notice
LinkedIn Ireland Unlimited Company	LinkedIn	Social network	45.2 (logged-in), 132.5 (logged-out site visits)	Privacy Policy European Regional Privacy Notice (EEA/UK) AI Features Privacy FAQ User Agreement
Meta Platforms Ireland Limited (MPIL)	Facebook	Social network	259	Privacy Policy Instagram Privacy Policy How Generative AI Models Work Generative AI Privacy Information Terms of Service EU AI Terms UK AI Terms Instagram Terms of Use
	Instagram	Social network	259	
Microsoft Ireland Operations Limited	Bing	Search engine	119	Privacy Statement Services Agreement Privacy FAQ for Microsoft Copilot
NKL Associates s.r.o	XNXX	Pornography video-sharing	45	Privacy Policy Terms of Service
Pinterest Europe Ltd.	Pinterest	Social network	124	Privacy Policy AI at Pinterest Terms of Service
Snap B.V.	Snapchat	Social network	102	Privacy Policy EEA/UK Privacy Notice Privacy by Product Perplexity on Snapchat Generative AI on Snapchat Terms of Service
TikTok Technology Limited	TikTok	Video-sharing social network	135.9	Privacy Policy (EEA) Terms of Service (EEA)

Twitter International Unlimited Company (TIUC)	X	Social network	115.1	Privacy Policy Data Processing Legal Bases Terms of Service
Whaleco Technology Limited	Temu	Retail platform	75	Privacy and Cookie Policy France Terms of Use Privacy and Cookie Policy
WebGroup Czech Republic	XVideos	Pornography video-sharing	160	Privacy Notice Privacy Policy Terms of Service
WhatsApp Ireland Ltd.	WhatsApp	Messaging platform	51.7	WhatsApp Privacy Policy (UK) WhatsApp Privacy Policy (EEA) WhatsApp DMA Notice for Non-Users WhatsApp Terms of Service (UK) WhatsApp Terms of Service (EEA)
Wikimedia Foundation Inc	Wikipedia	Online Encyclopaedia	151.1	Privacy Policy Terms of Use
Zalando SE	Zalando	Retail platform	74.5 (retail), 26.8 (third-party content)	Legal Notice Privacy Policy
Data source: https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses . Average monthly active users are as reported by providers to the European Commission. “Stripchat”, provided by Technius Ltd. is excluded as it is no longer designated.				

Public Policy Master's Thesis Series

This series presents the Master's theses in Public Policy and in European Affairs of the Sciences Po School of Public Affairs. It aims to promote high-standard research master's theses, relying on interdisciplinary analyses and leading to evidence-based policy recommendations.

Method in the Madness: Conceptualisations of User-Generated Platform Data in the Age of AI

Aron Thatcher West

Abstract

The processing of user-generated data from online platforms for the purposes of developing and using artificial intelligence systems has become widespread. This thesis examines how user-generated platform data and its processing in the context of AI model development are conceptualised and regulated under EU law and in platform policies. The EU legal framework for data is examined alongside the platform policies of 20 major providers. Three principal conceptualisations of data are identified based on relation to a person, availability, and economic characteristics. Whereas conceptualisations of data based on relation to a person and availability are common, an economic conceptualisation is found to have been under-theorised and under-implemented in law and platform policies, hindering the regulation of user-generated data and the effects such data can engender. A strengthened conceptualisation of data based on its principal economic characteristics is analysed, which it is argued would provide for a powerful means to complement the regulation of data processing in the context of AI model development, whilst opening the door to a range of wider public policies.

Key Words

Data, platforms, artificial intelligence, processing, law, economics.