Technology and Regulation

Harmed While Anonymous

Beyond the Personal/Non-Personal Distinction in Data Governance

Przemysław Pałka

data analytics;
data protection law; personal data;
non-personal data;
targeted advertising;
discrimination;
manipulation;
mental health

przemyslaw1.palka@uj.edu.pl

Data law and policy assume that harms to individuals can result only from personal data processing. Conversely, generation and use of non-personal data supposedly create new value while presenting no risk to individual interests or fundamental rights. Consequently, the law treats these two categories differently, constraining generation, use, and sharing of the former while incentivizing the latter. This article challenges this assumption. It proposes to divide data-related harms into two high-level categories: unwanted disclosure and detrimental use. It demonstrates how personal/non-personal data distinction prevents unwanted disclosure but fails to capture, and unintendedly enables, detrimental use of data. As a remedy, the article proposes a new concept – data about humans – and illustrates how it could advance data law and policy.

1. Introduction

The distinction between personal and non-personal data lies at the core of data regulations worldwide.[1] Law and policy approach these two categories very differently. Personal data is seen through the lens of fundamental rights,[2] and the law limits how much of it can be produced, how it can be used, or with whom it can be shared.[3] Non-personal data, on the other hand, is seen as an economic resource, which should be generated, used, and shared as much as possible.[4] The logic appears simple: data should be generated and used to produce value *unless* this data concerns an identifiable individual, in which case her rights trump the economic interests of market actors and society.[5] This approach, currently pursued and promoted by the European Union, assumes that threats to individuals result only from personal data processing.

This article draws attention to the fact that non-personal data about humans can be used to harm specific individuals and social interests. The claim is *not* that non-personal, anonymized data can easily be de-anonymized and turned (back) into personal data, as has been shown extensively in policy literature[6] and technical experiments.[7]

[1] Paul M Schwartz and Daniel J Solove, 'The PII Problem: Privacy and a New Concept of Personally Identifiable Information' (2011) 86 New York University Law Review 1814, 1816; 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, 43; Nadezhda Purtova, 'From Knowing by Name to Targeting: The Meaning of Identification under the GDPR' (2022) 12 International Data Privacy Law 163, 163; Maria Lilla Montagnani and Mark Verstraete, 'What Makes Data Personal?' (2022) 56 UC Davis Law Review 1165, 1169.

[2] Gloria González Fuster, The Emergence of Personal Data Protection as a Fundamental Right of the EU, vol 16 (Springer International Publishing 2014) 253–272 .

[3] Maximilian von Grafenstein, The Principle of Purpose Limitation in Data Protection Laws: The Risk-Based Approach, Principles, and Private Standards as Elements for Regulating Innovation (Nomos 2018) 109–124; Karen Yeung and Lee A Bygrave, 'Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship' (2022) 16 Regulation & Governance 137, 137–140.

[4] Simonetta Vezzoso, 'The Dawn of Pro-Competition Data Regulation for Gatekeepers in the EU' (2021) 17 European Competition Journal 391, 395–399; Josef Drexl and others, 'Position Statement of the Max Planck Institute for Innovation and Competition of 25 May 2022 on the Commission's Proposal of 23 February 2022 for a Regulation on Harmonised Rules on Fair Access to and Use of Data (Data Act)' (25 May 2022) 10–118 <https://papers.ssrn.com/abstract=4136484>

[5] See "A European Strategy for Data", Brussels, February 19, 2020, COM(2020) 66 final, at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066, p. 1, "Data-driven innovation will bring enormous benefits for citizens (…) any personal data sharing in the EU will be subject to full compliance with the EU's strict data protection rules (…) At the same time, the increasing volume of non-personal industrial data and public data in Europe, combined with technological change in how the data is stored and processed, will constitute a potential source of growth and innovation that should be tapped."

[6] Schwartz and Solove (n 1) 1841–1845; Purtova, 'The Law of Everything' (n 1) 47–48.

[7] Hiroshi Yoshiura, 'Re-Identifying People from Anonymous Histories of Their Activities', 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST) (2019); Alexandros Bampoulidis and others, 'Practice and Challenges of (De-)Anonymisation for Data Sharing' in Fabiano Dalpiaz, Jelena Zdravkovic and Pericles Loucopoulos (eds),

Assistant Professor, Faculty of Law and Administration, Jagiellonian University

The research leading to these results has received funding from the Norwegian Financial Mechanism 2014-2021, project no. 2020/37/K/HS5/02769, titled "Private Law of Data: Concepts, Practices, Principles & Politics."

Received 9 Mar 2023, Accepted 4 Jul 2023, Published 15 Sep 2023

Instead, the article maintains that even in situations of perfect anonymity – assuming, for the sake of argument, that such situations are technically possible – non-personal data about humans could be used to harm persons and communities. Consequently, the regulatory approach built around the distinction between personal and non-personal data is ill-suited to advance the policy goal of incentivizing value creation while protecting individuals from data-related harms.

Data-related harms to individuals, this article posits, could be divided into two high-level categories: *unwanted disclosure* and *detrimental use*. The former occurs when information about an individual becomes accessible to others against that individual's will and can range from a singular communication to publicizing private information to a data breach or theft. The latter transpires when data already in control of some actor (e.g., an online platform) is used to inflict damage upon an individual, for example, through behavioral manipulation,[8] discrimination in access to goods, services, or employment,[9] price discrimination,[10] or negatively impacting that individual's mental health.[11] According to contemporary laws like the EU's General Data Protection Regulation,[12] defining their scope of application, i.e., "personal data processing," widely, as "*any operation* (...) performed on personal data (...) such as collection, (...) use, (...) dissemination (...) or destruction,"[13] both categories of harm are potentially of interest to data protection law. However, the law grants protection only regarding actions concerning personal data, i.e., "information relating to an identified or identifiable natural person."[14]

The problem is that the second category of data-related harms – detrimental use – can easily be inflicted without identifying an individual.[15] For example, to discriminate against someone based on their race, gender, or sexual orientation, an algorithm filtering content does not need to identify the individual; it only needs access to their race, gender, or sexual orientation data[16] or proxy-data about these characteristics.[17] The individual in question can remain perfectly anonymous, their name, address, or social security number unknown, and still discrimination can occur. Moreover, statistical data about humans can be used in a general manner, e.g., to re-design the interface of some services, or the timing of notifications, to make them more addictive and therefore harmful to users' mental health.[18]

The personal/non-personal distinction makes sense as long as unwanted disclosure is concerned. Indeed, publicizing a search history of an anonymous individual presents little danger to privacy,[19] whereas the same action concerning an identified or identifiable person could be disastrous. However, using the same search history to filter communications in a discriminatory or manipulative manner can be equally effective and equally harmful, regardless of whether the individual is identifiable or not.

Consequently, this article argues that scholars and policymakers should go beyond the personal/non-personal distinction when pondering regulating the *use* of data and supplement it with a category of "data about humans." This broader category would include any information concerning humans, regardless of whether they are identified or identifiable, if only such data can be used to predict and influence human behavior or affect the legal or factual position that individuals or communities find themselves in. Importantly, this concept does not need to be enacted into any binding law; rather, it is an intellectual tool better suited to consider potential regulatory interventions when advancing the goals of value-creation while simultaneously protecting individuals from harms. In this sense, the article situates itself in the "how should we think?" rather than "what should we do?" kind of scholarship. Moreover, it is less concerned with the (proper or strategic) interpretation of the existing laws and more with advancing a more nuanced conceptual framework for thinking about policymaking. Finally, the proposal presented here is not meant to suggest any amendements to data protection law in the strict sense (like the GDPR) but to inform further policymaking in the data law understood broadly.

Three caveats. First, the article analyzes the European Union data law and policy, while accounting for the American counterparts, though the trends discussed here are global. The EU, for good and for bad, actively promotes its data governance model worldwide,[20] either through the extraterritorial application of its laws or through the so-called "adequacy decisions,"[21] or simply by serving as a blueprint

Research Challenges in Information Science (Springer International Publishing 2020).

[8] Eliza Mik, 'The Erosion of Autonomy in Online Consumer Transactions' (2016) 8 Law, Innovation and Technology 1; Daniel Susser, Beate Roessler and Helen Nissenbaum, 'Online Manipulation: Hidden Influences in a Digital World' (2019) 4 Georgetown Law Technology Review 1.

[9] Latanya Sweeney, 'Discrimination in Online Ad Delivery' [2013] arXiv:1301.6822 [cs] <http://arxiv.org/abs/1301.6822> Muhammad Ali and others, 'Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes' (2019) 3 Proceedings of the ACM on Human-Computer Interaction 199:1; Raphaële Xenidis, 'Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience' (2020) 27 Maastricht Journal of European and Comparative Law 736.

[10] Ramsi A Woodcock, 'Big Data, Price Discrimination, and Antitrust' (2016) 68 Hastings Law Journal 1371; Oren Bar-Gill, 'Algorithmic Price Discrimination When Demand Is a Function of Both Preferences and (Mis)Perceptions' (2019) 86 University of Chicago Law Review <https://chicagounbound.uchicago.edu/uclrev/vol86/iss2/12>.

[11] Tim Wu, 'Blind Spot: The Attention Economy and the Law Symposium: Innovative Antitrust' (2018) 82 Antitrust Law Journal 771; Allison Zakon, 'Optimized for Addiction: Extending Product Liability Concepts to Defectively Designed Social Media Algorithms and Overcoming the Communications Decency Act' (2020) 2020 Wisconsin Law Review 1107.

[12] Regulation (EU) 2016/679, of the European Parliament and 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 O.J. (L 119) [hereinafter GDPR].

[13] GDPR, art. 4.2, emphasis added.

[14] GDPR, art. 4.1.

[15] Though the matter might become more complicated if one adopts a broaded interpretation of "identification;" the problem is disussed in the following section.

[16] Sweeney (n 9); Ali and others (n 9).

[17] Anya ER Prince and Daniel Schwarcz, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data' (2019) 105 Iowa Law Review 1257.

[18] Zakon (n 11) 1113–1117; James Niels Rosenquist, Fiona M Scott Morton and Samuel N Weinstein, 'Addictive Technology and Its Implications for Antitrust Enforcement' (2021) 100 North Carolina Law Review 431, 433–435.

[19] And if it does, this is because of the possibility of identifying someone based on their search history. The problem exists only as long as it is possible to link the search history to a specific individual. A completely random search history (say, lacking names or addresses) would not be harmful to anyone even if made public.

[20] Anu Bradford, The Brussels Effect: How the European Union Rules the World (Oxford University Press 2020).

[21] European Commission, Adequacy decisions: How the EU determines if a non-EU country has an adequate level of data protection, at https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en. Countries that have received such a decision, i.e. aligned the logic of their data laws with the GDPR, include Argentina, Canada, Israel, Japan, New Zealand, Republic of

for other countries looking for models of data governance. Hence, the regulatory philosophy advanced by Brussels impacts data governance far beyond the Union's borders. This article does not claim that the specificities of data laws in other jurisdictions can be reduced to the skeleton of European law. Instead, it critically scrutinizes the logic globally promoted by the EU. Second, this article focuses solely on the corporate use of data in the private sector. This is not to minimize the significance of harms stemming from data use by public bodies. Rather, the objectives of data use by governments and corporations are sufficiently different, and the principles governing such uses are divergent enough for the separate analysis of the two to make sense. Finally, this article uses the terms "data" and "information" interchangeably. This mirrors the approach of the European and American legislators, who tend to define "X data as information Y." The relations between these two concepts are treated differently in various areas of study.[22] The article does not claim that these two notions are always equivalent; it simply bypasses the need to distinguish them by adopting the approach present in the positive law.

The article consists of three parts. First, it reconstructs the legal meaning of the concepts of "personal-" and "non-personal data" and analyzes the divergent philosophies of personal and non-personal data law. Second, it provides an overview of the technological and economic foundations of corporate data analytics, surveys the possible harms to individuals and groups, and demonstrates how these harms can materialize without processing personal data. Third, it introduces the concept of "data about humans" and discusses how this concept can inform the work of scholars and policymakers interested in data governance.

2. Data Regulation? Make it Personal (or not)

Data governance laws – the totality of norms stipulating who can perform what actions, upon what data, and under what conditions[23] – consist of various, often inconsistent, specific regulations.[24] The core distinction structuring data law and policy is between personal non-personal data.[25] When dealing with personal data, data protection and privacy laws apply; on the contrary, when dealing with non-personal data, they do not.[26] How does the law understand these terms?

2.2 What is (not) Personal Data?

A global definitional consensus is emerging: "personal data," or "personal information," should be understood as information concerning an *identified* or *identifiable individual*. Non-personal data, on the other hand, tends to be defined negatively as information that does *not* concern an identified or identifiable person. This latter category includes data about the environment or industrial data (unless linked to an identified or identifiable individual), but also data about humans that cannot be identified, like anonymous data about individuals or statistical data about groups.[27]

The GDPR defines "personal data" as:

> *Any* information *relating to* an *identified* or *identifiable* natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular 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.[28]

Not the content of information but its relation to a specific human being decides whether it will be considered "personal data."[29] Every piece of information, no matter how seemingly insignificant or important, can be regarded as both personal and non-personal data, depending on whether it relates to a natural person that is identified or identifiable.[30] Now, the notion of "identification" is less clear than could seem on its face, and has been elaborated upon by scholars, some of them arguing for more inclusive interpretations of the concept.[31] Acknowledging the fact that the line between an identified and a non-identified individual, and so consequently between personal and non-personal data, will always be context-specific and sometimes a good-faith disagreement about the law's interpretation can take place, this article does not focus on the myriad of possible interpretative problems, referring the interested reader to the excellent works doing so. Rather, this article focuses on the ideal types of personal and non-personal data, as currently imagined by data governance law and policy, and the consequences of the law treating them as one or the other. With this in mind, consider three examples:

E$_1$: There exists an individual who lives in Rome IT, is between 25 and 35 years old, listens to rap, is vegan, follows liberal politicians on Twitter, and identifies as gay and Catholic.

E$_2$: Giovanni dell'Esempio, social security number *******89, lives in Rome IT.

E$_3$: Persons who live in Rome IT, between age 25-35, who follow liberal politicians on Twitter, have a X% of chance being vegan, Y% chance of being gay, and Z% chance of being Catholic.

Korea, Switzerland, the United Kingdom, and Uruguay.

[22] Raphaël Gellert, 'Comparing Definitions of Data and Information in Data Protection Law and Machine Learning: A Useful Way Forward to Meaningfully Regulate Algorithms?' (2022) 16 Regulation & Governance 156.

[23] Salome Viljoen, 'A Relational Theory of Data Governance' (2021) 131 Yale Law Journal 573, 577; Przemysław Pałka, 'Data Management Law for the 2020s: The Lost Origins and the New Needs' (2020) 68 Buffalo Law Review 559, 566.

[24] Thomas Streinz, 'The Evolution of European Data Law' in Paul Craig and Gráinne de Búrca (eds), The Evolution of EU Law (3rd edn, 2021).

[25] Schwartz and Solove (n 1) 1816; Purtova, 'The Law of Everything' (n 1) 43; Purtova, 'From Knowing…' (n 1) 163; Montagnani and Verstraete (n 1) 1169.

[26] Lee A Bygrave and Luca Tosoni, 'Article 4(1). Personal Data' in Christopher Kuner and others (eds), The EU General Data Protection Regulation (GDPR): A Commentary (Oxford University Press 2020) <https://doi.org/10.1093/oso/9780198826491.003.0007>.

[27] Importantly, however, it is not the content of information but the relationship to the person that decides whether data will be treated as personal or not. Data about machinery or weather could be personal, if it related to and identified or identifiable individual. For an in-depth discussion of the problem, see: Montagnani and Verstraete (n 1).

[28] GDPR, art. 4.1., emphasis added, punctuation original. For the semi-authoritative elaboration of the concept, see Article 29 Data Protection Working Party Opinion 4/2007 on the concept of personal data, 01248/07/EN WP 136.

[29] Montagnani and Verstraete (n 1).

[30] Purtova, 'The Law of Everything' (n 1); Bygrave and Tosoni (n 26).

[31] Ronald Leenes, 'Do They Know Me? Deconstructing Identifiability' (2007) 4 University of Ottawa Law and Technology Journal 135; Frederik J Zuiderveen Borgesius, 'Singling out People without Knowing Their Names – Behavioural Targeting, Pseudonymous Data, and the New Data Protection Regulation' (2016) 32 Computer Law & Security Review 256; Michèle 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 11; Purtova, 'From Knowing…' (n 1).

E_1 does not – on its face – contain personal information. However, it includes a lot of anonymous information about a person, sensitive information for that matter (concerning their political opinions, sexual orientation, and religious affiliation). Still, this information does not necessarily qualify as "personal data" because the individual is not identified or identifiable. [32] Conversely, E_2 contains personal data as it relates to an identified individual. It does not matter that this information is generic and austere; what matters is that it relates to Giovanni dell'Esempio, a specific individual. E_3 definitely does not contain personal information as it only reports statistical data, albeit potentially very useful for marketers or other companies.

Again, the trickiest notion in the GDPR's definition is that of an "identifiable" natural person. As Nadezhda Purtova argues, under the broad concept of "identifiability," any dataset pertaining to a single person could be deemed as identifying that person.[33] Put simply: if there is only one human being in Rome that fits the description from E_1, this sentence contains personal data. However, the GDPR foresees situations where data concerning humans is not considered personal data, namely "anonymous information." Recital 26 of the GDPR states:

> The principles of data protection should therefore *not* apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to *personal data rendered anonymous* in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.[34]

The European Union has, for over a decade now, seen personal data protection as a core element of its constitutional identity.[35] More recently, however, the EU has expressed interest in advancing its technological capabilities in data analytics.[36] A range of regulations aimed at incentivizing data creation and sharing has been proposed or passed. Some of these regulations apply only to non-personal data;[37] others, like the Data Governance Act[38] or the Proposal for the Data Act,[39] to both personal and non-personal data. However, they all define non-personal data in the same way: data other than personal data according to the GDPR.

In the United States, the applicability of privacy laws has for a long time turned on whether one was dealing with "personally identifiable information" or the "PII."[40] Unlike the EU, the US has not adopted a horizontal federal statute governing data processing.[41] Consequently, American law lacks a uniform, commonly accepted definition of personally identifiable information, with different statutes and decision-makers approaching the problem differently.[42]

Recently, however, one can observe a move towards aligning the definitions of PII with the "identified of identifiable" model. On the state level, the 2018 California Consumer Protection Act (CCPA)[43] defines "personal information" as:

> Information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household.[44]

Similar approaches to defining personal data – focusing one being linked or linkable with an identified or identifiable individual – have been adopted by the lawmakers in Colorado,[45] Connecticut,[46] and Virginia,[47] and so the trend seems stable. On the federal level, the latest bipartisan proposal has been the American Data Privacy and Protection Act.[48] If passed, the ADPPA would be a revolution for the American data law. Albeit limited to the market context, it would be the first close-to-horizontal data protection bill applicable to all businesses. The bill is still under debate, and one should remember that the US has been trying to pass federal privacy legislation for the past two decades with no success.[49] However, if the ADPPA becomes the law, it would bring the American regime closer to the European model. The bill defines "covered data" as:

> Information that *identifies* or is *linked* or *reasonably linkable*, alone or in combination with other information, to an individual or a device that identifies or is linked or reasonably linkable to an individual may include derived data and unique persistent identifiers.[50]

Immediately, however, several exclusions are introduced, and those include "de-identified data," understood as:

> Information that does not identify and is not linked or reasonably linkable to an individual (...) regardless of whether the information is aggregated, provided that the covered entity takes reasonable technical, administrative, and physical measures to ensure that the information cannot, at any point, be used to re-identify any individual or device (...).[51]

[32] This could, in certain contexts, be personal information; however, simply as a sentence printed on this page, E_1 does not qualify as personal information.

[33] Purtova, 'The Law of Everything' (n 1) 46.

[34] GDPR, recital 26, emphasis added.

[35] Bilyana Petkova, 'Privacy as Europe's First Amendment' (2019) 25 European Law Journal 140.

[36] European Strategy for Data (n 5).

[37] 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, 2018 O.J. (L 303)

[38] The 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), 2022 O.J. (L152/1).

[39] Proposal for a Regulation on harmonised rules on fair access to and use of data (Data Act), COM/2022/68 final.

[40] Schwartz and Solove (n 1).

[41] Which does not, however, mean that there is no federal privacy law at all; some scholars treat the Federal Trade Commission's jurisprudence as one, see Daniel J Solove and Woodrow Hartzog, 'The FTC and the New Common Law of Privacy' (2014) 114 Columbia Law Review 583.

[42] Schwartz and Solove (n 1).

[43] California Consumer Protection Act, CAL. CIV. CODE §§ 1798.100-.192 [hereinafter CCPA]

[44] Id. sec. 1798.140.0.(1).

[45] Colorado Privacy Act, Senate Bill 21-190, sec. 6-1-1303.(17).

[46] An Act Concerning Personal Data Privacy and Online Monitoring, Public Act No. 22-15, Sec.1(18).

[47] Code of Virginia, Chapter 53. Consumer Data Protection Act, § 59.1-575.

[48] American Data Privacy and Protection Act Bill H.R.8152, introduced in House 06/21/2022. [hereinafter ADPPA]

[49] Jessica Rich, 'After 20 Years of Debate, It's Time for Congress to Finally Pass a Baseline Privacy Law' (Brookings, 14 January 2021) <https://www.brookings.edu/blog/techtank/2021/01/14/after-20-years-of-debate-its-time-for-congress-to-finally-pass-a-baseline-privacy-law/>.

[50] ADPPA. sec. 2.8.A.

[51] Id. sec. 2.10.

Consequently, one can observe precisely the same move made in the ADPPA that the EU has taken in the GDPR: first, personal data is defined broadly as any information relating to an identified or identifiable individual. Second, non-personal data, i.e., the frontier of the law's applicability, is understood as any information not relating to an identified or identifiable individual. Regardless of whether this non-personal data concerns weather, machinery, or flesh-and-bones human beings, if only anonymous.

What are the normative, both legal and policy, consequences of considering a piece of information personal or non-personal data?

2.2 Different Logics of Personal and Non-Personal Data Laws

The logic of personal and non-personal data laws and policies is almost perfectly opposite.

Non-personal data is treated as a non-rivalrous resource that can and should be used to generate value.[52] Policymakers have deployed various strategies to ensure that data about the market is available to as broad an audience as possible. First, there are numerous regulations requiring sharing information, ranging from various mandated disclosures to consumers[53] – or, in the European legal jargon, "information obligations"[54] – to reporting obligations in the capital markets. Second, when companies share information on their own motion, through advertising or other commercial practices, consumer law forbids communications (and omissions) that are untrue, misleading, or deceptive.[55] Third, the law has taken steps against ownership of information, either explicitly in IP law or implicitly, by not granting such rights through other means.[56] Fourth, under long-standing European regulations, publicly generated data should be freely accessible to all and reusable for commercial and non-commercial purposes.[57] All this results from a conviction deeply rooted in the market logic: for the economy to be efficient, data about producers, consumers, and available goods and means should be abundant, easy to access and use for all.[58]

Most recent legislative activity in the European Union takes this logic further. In 2018 the EU adopted a Regulation on Free-Flow of Non-Personal Data to remove obstacles to data sharing within the Union. Realizing that its impact has been limited, in May 2022, the EU adopted the Data Governance Act, stating in its opening recitals that:

> Over the last decade, digital technologies have transformed the economy and society (...). Data is at the centre of that transformation: *data-driven innovation will bring enormous benefits* to both Union citizens and the economy. (...) It is necessary to improve the conditions for data sharing in the internal market (...). This Regulation should aim to *develop further* the borderless digital internal market and a human-centric, trustworthy and secure *data society and economy*.[59]

Further, the Digital Markets Act[60] imposes several data-sharing-related obligations on the so-called "gatekeepers," including data interoperability and access for competitors.[61] Finally, the still-under-debate Data Act aims to free the data generated by Internet of Things devices. The explanatory memorandum opens by stating that:

> Data is a core component of the digital economy, and an essential *resource* to secure the green and digital transitions. The volume of data generated by humans and machines has been increasing exponentially in recent years. *Most data are unused however*, or its value is concentrated in the hands of relatively few large companies. (...). It is therefore crucial *to unlock such potential* by providing *opportunities for the reuse of data*, as well as by removing barriers to the development of the European data economy (...).[62]

Leaving aside the question of whether these instruments will achieve the stated goals, the logic is clear. In the recent years, the amount of generated data has grown exponentially, which is a good thing (according to the regulations). What the European lawmaker considers the problem is that data remains unused and that access to data, and the benefits of its use, are limited to a small number of actors. Hence, data should be easier to access and use, and the opportunity for, as well as benefits of such use, should be shared more widely.

Consequently, if one wanted to spell out the general principles of the European non-personal data law, these would include:

(i) data maximization (the more data generated, the better),
(ii) freedom of use (subject to rare prohibitions, companies are free to use non-personal data for whatever purpose they wish, the more innovative the use, the better),
(iii) access (non-personal data should be accessible to as wide a range of actors as possible),
(iv) public good (non-personal data belongs to no one and should be used to advance the public interest).

How do these principles fare when compared to the logic of personal data law, exemplified by the GDPR?

[52] Vezzoso (n 4); Drexl and others (n 4).

[53] Omri Ben-Shahar and Carl E Schneider, More Than You Wanted to Know: The Failure of Mandated Disclosure (Princeton University Press 2014).

[54] Gert Straetmans, 'Information Obligations and Disinformation of Consumers' in Gert Straetmans (ed), Information Obligations and Disinformation of Consumers (Springer International Publishing 2019) .

[55] Willem van Boom and Amandine Garde, The European Unfair Commercial Practices Directive: Impact, Enforcement Strategies and National Legal Systems (Routledge 2016); Luke Herrine, 'The Folklore of Unfairness' (2021) 96 New York University Law Review 431.

[56] P Bernt Hugenholtz, 'Against "Data Property"' [2018] Kritika: Essays on Intellectual Property <https://www.elgaronline.com/view/edcoll/9781788971157/9781788971157.00010.xml> accessed 25 May 2020; Ignacio Cofone, 'Beyond Data Ownership' (2021) 43 Cardozo Law Review 501.

[57] Sara Gobbato, 'Open Science and the Reuse of Publicly Funded Research Data in the New Directive (EU) 2019/1024' (2020) 2 Journal of Ethics and Legal Technologies 145.

[58] Kenneth J Arrow and Gerard Debreu, 'Existence of an Equilibrium for a Competitive Economy' (1954) 22 Econometrica 265; Luke Herrine, 'What Is Consumer Protection For?' (2022) 33 Loyola Consumer Law Review <https://papers.ssrn.com/abstract=3781762>.

[59] Data Act Proposal (39), recitals 2 and 3, emphasis added, British spelling original.

[60] 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), 2022 O.J. (L 265).

[61] Vezzoso (n 4).

[62] Data Act, emphasis added.

	Non-personal data	Personal data
Generation	Good	Suspicious
Use	Good	Suspicious
Purpose of use	Anything not illegal	Only the agreed upon
Sharing	Good	Suspicious
Access	Unlimited	Limited
Individual vs. collective	Common good	Individual rights
Transparency	No	Yes
Principles	Data maximization, freedom of use, access, public good	Data minimization, purpose limitation, secrecy, individual rights

Table 1 Data law and policy's view on non-personal and personal data

First, the fundamental principle of the GDPR is purpose limitation.[63] Personal data can be used only for the purpose for which it was collected. Second, the GDPR introduces the principles of data minimization and storage limitation.[64] No more data than necessary to achieve the stated purpose can be collected, and it cannot be stored longer than necessary to realize that purpose. Third, the GPDR makes it difficult to share personal data with third parties. If an online platform uses the services of some other company (for example, to monitor web traffic), it must sign a data processor contract and inform the data subjects.[65] Fourth, the GDPR grants the data subjects a range of rights, including the right to object to processing, to withdraw consent at any time, to be forgotten, etc.[66] Finally, the GDPR imposes strict transparency obligations on entities processing personal data, including consumer-oriented disclosures through privacy policies[67] and regulator-oriented accountability obligations.[68]

Hence, the logic of personal and non-personal data law is precisely the opposite. Non-personal data generation is seen as, in principle, good, whereas personal data generation is always suspicious. Non-personal data can be used for any purpose not forbidden by the law, while personal data only for the purpose agreed upon by the data subject. Non-personal data should be shared and accessible widely, but personal data should be kept secret and secure. Non-personal data is a common good that everyone can benefit from, whereas personal data is, first and foremost, the domain of the data subject. Personal data processing must be transparent, whereas no such requirements exist for non-personal data processing. This is succinctly illustrated in Table 1.

This all looks good on paper, but the reality proves more complicated. Many of the datasets include personal and non-personal data mixes, and the line between the two – given the broad definition of the "identifiable" individual – is not always clear.[69] However, for the sake of this paper's argument, let us assume that the distinction holds, i.e., it is possible to render data about humans anonymous to a degree in which it can no longer be linked to a specific individual. Why would one assume that, given the good amount of evidence to the contrary?[70]

It is in the economic interest of corporations to process non-personal data rather than personal data if only the same goals can be achieved in this way. Compliance with the GDPR is costly, and the potential fines for its violation are high.[71] Moreover, as demonstrated above, one is free to use non-personal data for many more purposes, in a much less constraining legal environment. Hence, given the innovative abilities of corporations like Google, Meta, or Amazon, and in the light of the clear policy message – "process as little personal data as possible, and as much non-personal data as possible" – one can assume that a significant amount of resources will be devoted to the anonymization of data. At least for as long as non-personal data can be used to achieve the same commercial goals. Further, a growing amount of scientific evidence demonstrates that it is possible to create anonymous protocols for the communication of IoT devices[72] and even anonymous online IDs.[73] The matter is not technologically trivial, but neither is it impossible to accomplish.

Under the current narrative of the European Union, such a world – a world where very little personal data is processed while non-personal data is abundant – would be close to a dream come true. If only this non-personal data is further widely accessible to competitors, the market forces will deliver a quickly growing, innovative economy, in which the rights of individuals are not in jeopardy. Or so the narrative goes.

European lawmakers assume that data-related harms to individuals can only be inflicted when processing personal data. However, this is not the case.

[63] von Grafenstein (n 3).
[64] GDPR, art. 5.1.c and 5.1.e.
[65] GDPR. art. 28.
[66] GDPR Chapter III.
[67] GDPR art. 12-14.
[68] GDPR art. 30.
[69] Purtova, 'The Law of Everything' (n 1).
[70] Yoshiura (n 7); Bampoulidis and others (n 7).
[71] GDPR art. 83.
[72] Jayasree Sengupta, Sushmita Ruj and Sipra Das Bit, 'End to End Secure Anonymous Communication for Secure Directed Diffusion in IoT', Proceedings of the 20th International Conference on Distributed Computing and Networking (Association for Computing Machinery 2019) <https://doi.org/10.1145/3288599.3295577>.
[73] Ray Kresman, Larry Dunning and Jinglei Lu, 'An Improved Anonymous Identifier', 2022 10th International Symposium on Digital Forensics and Security (ISDFS) (2022).

3. Data Economy: Mechanics, Opportunities, and Risks

This article proposes to divide data-related harms into two high-level categories: *unwanted disclosure* and *detrimental use*. The former occurs when a company publicizes information related to an identified or identifiable individual without their consent. For such harm to be possible, the company needs to control personal data in the first place; mere disclosure of data that cannot be linked to any specific individual does not harm any specific individual, by definition.

Detrimental use, on the other hand, occurs when a company uses data about humans in a manner harmful to an individual or a social group. This can happen when data is used to discriminate against members of protected groups,[74] extract consumer surplus through price discrimination,[75] manipulate individuals' preferences or behavior,[76] or inflict harm upon their mental health.[77] Crucially, for these actions to be possible, a company does *not* need access to personal data – data linked or linkable to a specific individual – all it needs is anonymous data about individuals and statistical data about humans.

These two categories of harm neatly track the distinction introduced by Raphaël Gellert, who argues that the GDPR is grounded in the logic of knowledge communication, whereas machine learning (a technology underpinning many types of detrimental use) in the logic of knowledge generation.[78] Indeed, harms potentially stemming from communication are well-addressed by the GDPR. However, those resulting from insight application – a different logic indeed – escape the Regulation's reach.

To better understand how non-personal data about humans can be used to inflict harm upon individuals and groups, one needs to analyze three questions: first, how and why do corporations use data about humans (what technologies do they deploy, and what economic incentives do they act upon)? Second, what risks to individuals are associated with these uses? Third, to what extent does the link between data and an identifiable individual matter for the corporate benefits and individual harms? Let us address them in this order.

3.1 The Underlying Technologies and the Economic Incentives

Data is valuable for corporations when it can serve as a source of actionable knowledge advancing the company's goals.[79] Turning data into operationalizable insights is possible due to advances in data science, most importantly machine learning. These advances took place not only due to the improvement of algorithms and increased computing power but also due to the sudden availability of large amounts of data to be analyzed.[80]

The sociotechnological precondition for the emergence of the data economy has been the gradual digitalization of everyday life. A growing number of socioeconomic activities – from communication and knowledge acquisition, over commerce and transportation, to entertainment and dating – have become mediated by technology. Online platforms and mobile apps like Facebook, Google, Amazon, Uber, or Tinder engage in "datafication," i.e., keeping digital records of an ever-growing number of human activities.[81]

Data can be monetized in many ways, but among the most profitable ones is programmatic advertising.[82] The "service for data" business model has proven so profitable that some internet giants, like Meta or Alphabet, earn billions of dollars per quarter without asking individual consumers for any monetary compensation.[83] Hence, advertising is a gentle case study to introduce the mechanics of modern data analytics.

The revenue generated by products like Facebook or Google is directly proportional to two metrics: the number of ads displayed and the price of each individual ad placement. To increase the price, corporations develop ad-delivery systems appealing to advertisers, i.e., effective in consumer preference and behavior modification (growth in sales). Meta and Alphabet will therefore aim at showing particular ads to the consumers with the highest probability of purchasing the product, in a mode further increasing the probability. They will, therefore, attempt to personalize not only the content of an ad (say, a grooming kit for a person thinking of starting to grow a beard) but also the form (what photo will the ad feature?), timing (morning, afternoon, workday, weekend?) and context (will the ad appear on the feed after a happy photo of a friend or an outrageous article published by the New York Times?), as all these elements play a role in the effectiveness of advertising.[84]

Consequently, the task that Meta and Alphabet face is to establish what persons are the most prone to purchase a particular product and what timing, form, and context of the ad will further increase that chance.[85] In principle, the task is no different from what advertising agencies have been doing long before the emergence of the digital economy.[86] What differentiates the tech giants from their analog predecessors is that they can automate the task by relying on data science and machine learning.

Ethem Alpaydin explains the essence of the process, stating:

> Once, it used to be the programmer who defined what the computer had to do, by coding an algorithm in a programming language. Now for some tasks, we do not write programs but collect data. The data contains instances of what is to be done, and the learning algorithm modifies a learner program automatically in such a way so as to match

[74] Sweeney (n 9); Ali and others (n 9); Xenidis (n 9).
[75] Woodcock (n 10); Bar-Gill (n 10).
[76] Mik (n 8); Susser, Roessler and Nissenbaum (n 8); Julie E Cohen, Between Truth and Power: The Legal Constructions of Informational Capitalism (Oxford University Press 2019).
[77] Wu (n 11); Zakon (n 11).
[78] Gellert (n 21).
[79] Laura Igual and Santi Seguí, Introduction to Data Science (Springer International Publishing 2017) <http://link.springer.com/10.1007/978-3-319-50017-1>.
[80] Ethem Alpaydin, Machine Learning: The New AI (The MIT Press 2016).
[81] Cohen (n 76).
[82] Tim Hwang, Subprime Attention Crisis: Advertising and the Time Bomb at the Heart of the Internet (FSG Originals 2020).
[83] In the first quarter of 2022, Meta Inc. reported the revenue of $27,908,000,000 (over twenty-seven billion dollars), 97% of which came from advertising, source https://investor.fb.com/investor-news/press-release-details/2022/Meta-Reports-First-Quarter-2022-Results/default.aspx; in the same period, Alphabet Inc., reported revenue of $ 68,011,000,000 (over sixty-eight billion dollars), out of which 80% came from advertising, source https://abc.xyz/investor/static/pdf/2022Q1_alphabet_earnings_release.pdf
[84] Mik (n 8).
[85] Mik (n 8).
[86] Tim Wu, The Attention Merchants: The Epic Scramble to Get Inside Our Heads (Vintage Books, a division of Penguin Random House LLC 2017).

> the requirements specified in the data.[87]

Hence, Alphabet or Meta will display ads to various people, regarding whom they already have vast amounts of data (what accounts they follow, what pages they visit, etc.), and observe who ends up clicking on what ads (generating new, feedback data). The result of the process will be a new data set, automatically annotated to record a detected pattern: when advertising product X, it should be displayed to individuals with Y characteristics in a Z manner.[88] This allows corporations to use data to increase the price metric; the better the ads, the higher the price.

Further, there is the metric of the number of ads displayed. Meta or Alphabet want people to spend as much time as possible using their products. This can be accomplished by personalizing the content displayed on users' feeds in a way that makes them keep scrolling or by properly timing the notifications in a way that makes users often come back.[89] As Meta already has large amounts of data about its users, it can play with various kinds of content displayed or notification patterns employed, and observe what methods lead to its desired outcomes. Availability of data allows corporations to rely on machine learning, whereas availability of machine learning incentivizes them to collect even more data.

Machine learning is just one method within the growing toolbox made available by the advances in data science. Laura Igual and Santi Seguí define it as:

> A methodology by which actionable insights can be inferred from data (…) the production of beliefs informed by data and to be used as the basis of decision-making.[90]

The application of various analytical tools to large datasets allows corporations to (i) discover patterns in human behavior and (ii) predict future events with a specific degree of probability.[91] This knowledge can then be (iii) put into action by algorithms, for example, those driving programmatic advertising[92] or recommender systems.[93] However, these tools can be used to automate any task that a specific corporation might want to optimize.

Amazon or Uber, with the ability to price-discriminate,[94] might want to increase the price for individuals with a higher willingness to pay and decrease it for others to maximize the revenue from the overall sales. Netflix or HBO Max might want to perfect their recommender systems to increase the customers' satisfaction and minimize the chance of them canceling subscriptions. LinkedIn or ZipRecruiter might want to improve their ability to match job seekers with potential employers to increase the chance that both will pay for their services. Each of these processes can be automated by data analytics, and the more (quality) data there is, the better the performance will be.

Importantly, for data analytics to work well, one needs large amounts of data,[95] what came to be known in law and policy circles as big data.[96] Hence, what matters is not individual data points regarding a specific person but large databases documenting *social* behavior, enabling an analyst to detect patterns. The operational paradigm of data-driven task automation is not certainty but probability. One can never know if Giovanni dell'Esempio will actually click on an ad, buy a product made more expensive just for him, or apply for a suggested job. However, one can know that statistically speaking, a person with Giovanni's characteristics has a specific, say 67%, chance of clicking, buying, or applying when displayed concrete content in a particular manner.

All these advancements benefit both corporations and, to some extent, the consumers. Higher revenues resulting from lower costs or increased profits translate into larger shareholder dividends and lower consumer prices. More data-driven innovation can lead to new, quality consumer products. However, as data policy scholars have pointed out, with the benefits to individuals and communities also come risks to their interests and fundamental rights.

3.2 Data-Related Harms and Why they Happen

Let us take a closer look at what harms can befall an individual because corporations control vast amounts of data. Imagine a company – like Meta or Alphabet – controls a lot of personal data about Giovanni dell'Esempio: his online browsing history, social media activity, the apps he has installed on his phone, etc. To make him more concrete, imagine Giovanni is indeed the person from E_1 in the previous section. In what ways could a company harm him?

First, it could *disclose* some of the information without Giovanni's consent. It could send a message to Giovanni's friends, telling them what websites he likes to visit. It could publicize Giovanni's search and browsing history on a local news website. Such actions would constitute intentional disclosure of private facts. Moreover, it could disclose data unintentionally, either by mistake or negligence in securing the data, followed by getting hacked by some nefarious actor. All such actions are, without a doubt, harmful to Giovanni's privacy and illegal not only under the data protection laws but also under general tort law.[97]

Why would any company do that? Except for very specific cases, usually politically loaded, corporations have no incentive to disclose their users' personal data without their consent. If using Facebook, Google, or Amazon, came with the risk of our personal data becoming available to others without our consent, people would be reluctant to use them. The corporate incentive is exactly the opposite: to keep the information users consider private, private. The story is slightly more complicated with cybersecurity – ensuring that data is safe and difficult to access without authorization is obviously costly – but there is a market incentive to invest in such precautions.

Unwanted disclosure is bad for business. The interests of Meta and Giovanni dell'Esempio, when it comes to keeping his personal data secret, perfectly align.

[87] Alpaydin (n 80) IX.
[88] This is obviously is a huge simplification; the result will be a large table with dozens of variables and values, expressed in probabilities. This sentence is supposed to capture the overall logic of the machine learning process.
[89] Zakon (n 11).
[90] Igual and Seguí (n 79) 2.
[91] Igual and Seguí (n 79) 2-3.
[92] Hwang (n 82).
[93] Igual and Seguí (n 79) 165–179.
[94] Bar-Gill (n 10).
[95] Alpaydin (n 80).
[96] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671; Woodcock (n 10); Prince and Schwarcz (n 16).
[97] William L Prosser, 'Privacy' (1960) 48 California Law Review 383.

Matters become more complicated with detrimental use. Here, the incentives of the user and the corporation might be inconsistent with one another. What is "detrimental" to the user might often be profitable for the corporation.

Consider the amount of time users spend on platforms like Instagram or Twitter. Their owners want to maximize it as it translates into more data generated and more ads displayed. Hence, they will show users content that encourages them to keep scrolling or time the notifications in a way that leads users to keep coming back or design the services in a way that encourages prolonged use.[98] All these processes are data-driven and can be automated. However, spending as much time as possible on Instagram or Twitter is not necessarily in the interest of consumers. A rapidly growing body of literature devoted to social media addiction[99] indicates that some users spend more time on these platforms than they want and engage in continued use despite negative consequences. Such excessive use is correlated with lower productivity,[100] anxiety and depression[101] or eating disorders.[102] First studies documenting the causal effect are being published.[103] Of course, one cannot claim that companies like Meta want their users to be addicted, depressed, or suffer from anorexia. All they want is for them to spend a lot of time using their services. They optimize for engagement, not problems for mental health. However, the negative mental health impacts are the (unintended) consequences of this optimization.

Further, consider the effectiveness of ads. As discussed in the previous subsection, Meta and Alphabet will fine-tune delivered ads to maximize their effectiveness.[104] However, it is not necessarily in the interest of consumers to buy all this stuff. Scholars studying the impact of behavioral advertising on consumer purchasing behavior point out that, given the sophistication of ad-personalization systems, the autonomy of consumers is in jeopardy.[105] Admittedly, it is difficult to establish the precise border distinguishing consumers acting upon their short-term preferences, which they later regret, and being manipulated into a purchase; the quantification of the scale of this phenomenon is therefore challenging. However, at least theoretically and anecdotally, it is easy to imagine or remember a person ordering something online when tired, stressed, or otherwise vulnerable[106] only to then ask oneself, "why would I ever have bought this?"

Ad-delivery can also have a discriminatory impact. Empirical studies demonstrate how postings for high-paid jobs tend to be displayed predominantly to white males.[107] This can happen either as a result of flawed datasets used for training the algorithm or, more dangerously, because the ad-delivery algorithm replicates the existing historical inequalities captured in the data.[108] Again, it would be a stretch to claim that Meta or Alphabet want to discriminate against anyone; they optimize for the chance of clicking. If an advertiser paid for 1000 ads to be delivered, and algorithms associate the potential interest with the characteristics of being white and male, it will show ads to people with these characteristics. Still, this can lead to disparate impact[109] or, in the European legal jargon, indirect discrimination.[110] Importantly, the algorithms do not need to work openly on the data points indicating "white" or "male," it can end up delivering the ads to such individuals based on proxy data, seemingly unrelated, like the kinds of pages one follows on Facebook or the kinds of websites one looks for on Google.[111]

One could keep going and discussing potential harms in detail. An individual shown a higher price because an algorithm assessed their willingness to pay as high would have, subjectively speaking, overpaid.[112] An individual shown specific search results, suboptimal from their point of view, but optimal from the point of view of the company's goals will have suffered harm to their autonomy or equality. However, the point of this section has been to demonstrate that the interests of corporations using data about humans, and the customers of their services, do not always align. Hence, the use of such data can be detrimental to individuals and social groups as a byproduct of a corporation pursuing its own interests.

The difference between the two kinds of data harms – unwanted disclosure and detrimental use – is double. First, the former is where the interests of corporations and consumers align, whereas the latter is where they often are contrary. Second, unwanted disclosure can only happen when processing personal data, whereas detrimental use can easily be a result of non-personal data processing. Let us see how.

3.3 No Need to Identify

When analyzing the process of deriving actionable insights from data, one should distinguish two phases: knowledge generation and knowledge application. In practice, these are not always completely separable but, logically speaking, it helps to analyze them this way. When it comes to knowledge application, one can further distinguish two modes: general and individual.

Knowledge generation always concerns data related to groups and aims to answer questions like: persons with what characteristics are most likely to click on an ad for such a product? What is the best interval to serve users notifications to maximize the amount of time

[98] Zakon (n 11).

[99] Qinghua He, Ofir Turel and Antoine Bechara, 'Brain Anatomy Alterations Associated with Social Networking Site (SNS) Addiction' (2017) 7 Scientific Reports 45064; Nazir S Hawi and Maya Samaha, 'The Relations Among Social Media Addiction, Self-Esteem, and Life Satisfaction in University Students' (2017) 35 Social Science Computer Review 576; Yubo Hou and others, 'Social Media Addiction: Its Impact, Mediation, and Intervention' (2019) 13 Cyberpsychology: Journal of Psychosocial Research on Cyberspace <https://cyberpsychology.eu/article/view/11562> Zakon (n 11).

[100] Cal Newport, Deep Work: Rules for Focused Success in a Distracted World (1st edition, Grand Central Publishing 2016).

[101] Amandeep Dhir and others, 'Online Social Media Fatigue and Psychological Wellbeing—A Study of Compulsive Use, Fear of Missing out, Fatigue, Anxiety and Depression' (2018) 40 International Journal of Information Management 141.

[102] Siân A McLean and others, 'Selfies and Social Media: Relationships between Self-Image Editing and Photo-Investment and Body Dissatisfaction and Dietary Restraint' (2015) 3 Journal of Eating Disorders O21.

[103] Melissa G Hunt and others, 'No More FOMO: Limiting Social Media Decreases Loneliness and Depression' (2018) 37 Journal of Social and Clinical Psychology 751.

[104] Mik (n 8).

[105] Mik (n 8); Jan Trzaskowski, Your Privacy Is Important to Us! – Restoring Human Dignity in Data-Driven Marketing (ExTuto 2021).

[106] Natali Helberger and others, 'EU Consumer Protection 2.0: Structural Asymmetries in Digital Consumer Markets.' (2021) <https://www.beuc.eu/publications/beuc-x-2021-018_eu_consumer_protection.0_0.pdf>.

[107] Sweeney (n 9); Ali and others (n 9).

[108] Barocas and Selbst (n 96).

[109] Barocas and Selbst (n 84).

[110] Xenidis (n 9).

[111] Prince and Schwarcz (n 16).

[112] Woodcock (n 10); Bar-Gill (n 10).

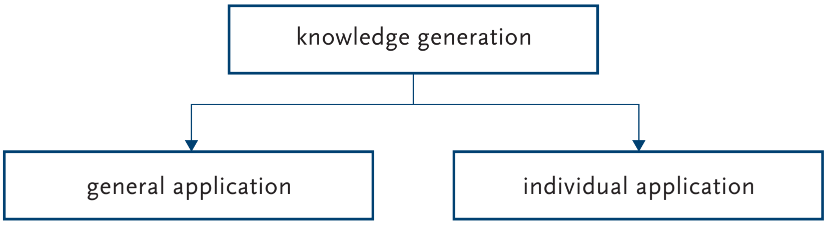

Figure 1 Phases of data-driven knowledge application

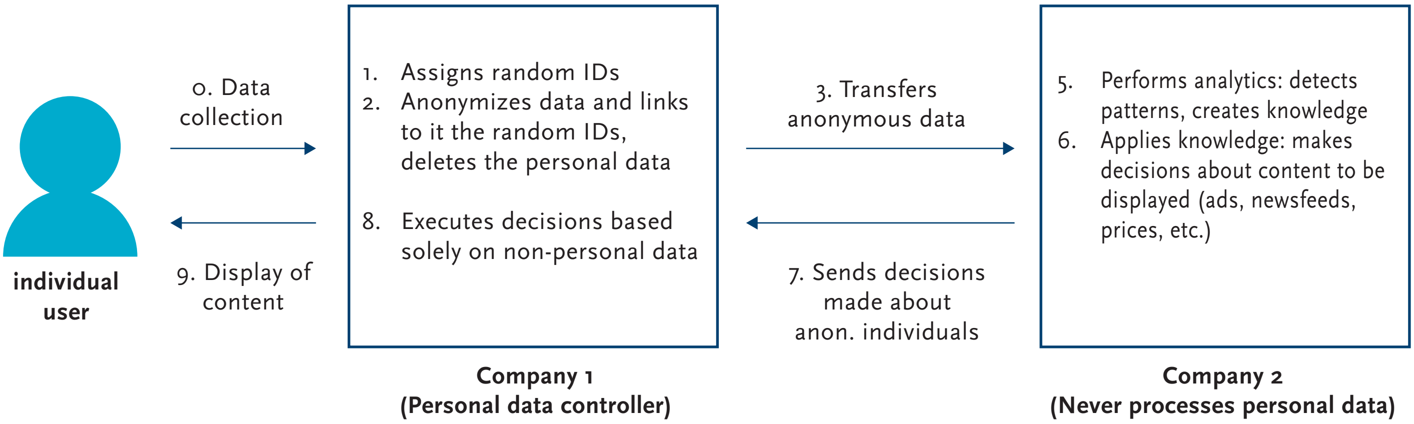
individual user
0. Data collection
9. Display of content
1. Assigns random IDs
2. Anonymizes data and links to it the random IDs, deletes the personal data
8. Executes decisions based solely on non-personal data
3. Transfers anonymous data
7. Sends decisions made about anon. individuals
5. Performs analytics: detects patterns, creates knowledge
6. Applies knowledge: makes decisions about content to be displayed (ads, newsfeeds, prices, etc.)
Company 1 (Personal data controller)
Company 2 (Never processes personal data)

Figure 2 The processes of applying insights based solely on non-personal data

they spend using the app? What characteristics are correlated with a higher willingness to pay?

Those are all statistical questions. To answer them, personal data is not necessary. What interests a company concerns scalable characteristics (age, domicile, interests, etc.) and related behavior (clicking, reacting, accepting a higher price, etc.). By stripping a database of all data enabling identification of an individual – names, phone numbers, IP addresses, etc. – one does not lose any information relevant to answering these questions. To generate knowledge from data, this data can be fully anonymous. General application of such knowledge also does not require processing of personal data. If a corporation discovers the optimal (from its point of view) interval of notifications, it can deploy it across the entire service. There is no connection between an identifiable person and the overall change of the service.

The matter becomes more technically complex when it comes to individual application of the insights generated from data. In the end, an algorithm will need to determine what ad to show Giovanni dell'Esempio (and what not), what to display on his newsfeed, what price to offer him for a particular service, etc. It is this moment that most often will fall within the scope of application of personal data law, as communicating content to a specific individual involves processing of data concerning an identified person. Hence, one could assume that the individual is protected throughout the entire cycle of data processing. Not necessarily.

Imagine Meta splits into two companies, one responsible for tasks inherently connected to processing of personal data – storing and processing of photos, account verification, etc. – and another engaged in data analytics, insights from which will be deployed for advertising, newsfeed curation, price discrimination, etc. The former company assigns each user a randomly-generated ID number and then associates observed behavior – what content a user reacts to, in what way, etc. – with that ID. It makes sure that any data that could identify a specific individual is removed. Once the anonymized data set is ready, it immediately deletes all the personal data collected in the first place. Then, it shares the generated data set – containing only non-personal data – with the second company which is responsible for dealing with advertisers. That other company makes decisions on what user will see and sends the order back to the first one. Hence, the entity able to link a particular ad, or newsfeed content, with a specific individual will have no idea why the user sees which ad and will not have processed any personal data while determining so.

Within such a structure, actions 0., 1., and 2. (and arguably, under the broad interpretation of "identification,"[113] also 8. and 9.) would fall within the scope of personal data law and be subject to principles like data minimization or purpose limitation. However, the "heavy-lifting" operations – pattern detection in data analytics, decisions on what to display whom, etc. – are performed upon non-personal data and therefore governed by an exactly opposite normative logic. One could argue that such a structure is designed to exploit the loophole in the GDPR. However, just as well, it could be seen as staying faithful to the letter and the spirit of the GDPR: minimizing the amount of personal data processed through anonymization. The important point is that to generate actionable insights from data – insights furthering corporate goals while potentially threatening the interests of individuals – one does not need to process personal data.

Given the advances in computer science, it might even be possible to run such operations anonymously within one company. Researchers

[113] Purtova, 'From Knowing...' (n 1).

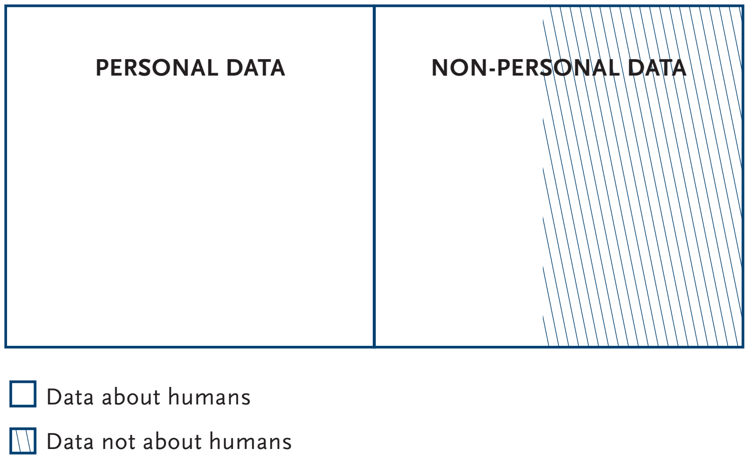

Figure 3 The relationship between personal/non-personal data to data (not) about humans

have demonstrated how by assigning randomized IDs to individuals, their personal data can be anonymized, shared with other nodes in the network, and acted upon again.[114] Similar tools are being developed for anonymous data sharing by IoT devices.[115] The matter is not technically trivial, but it is not impossible. And there are legal incentives to invest in these technologies and corporate processes further.

In this sense, individuals can suffer detrimental-use-of-data-related harms without having their personal data processed for most of the data-analytics cycle. This fact escapes the narrative of data law and policy, as of today, tacitly assuming that harms to individuals can materialize only when processing personal data. A correction is due.

4. Beyond the Personal/Non-Personal Distinction in Data Governance

To account for possible data-related harms to individuals and social groups, and potentially mitigate them, data law and policy need a new concept: "data about humans." The introduction of this concept can help nuance the binary logic assumed in data law and policy, according to which personal data processing is always suspicious while non-personal data processing always laudable. In fact, whenever data about humans is processed, there is a risk of harm.

Clearly, simply introducing a concept will not solve any real-world problems. Hence, the article concludes by discussing the potential goals and means of reforming the data laws in the EU and beyond. Each of discussed proposals has its strengths and weaknesses; the goal of this last section is not to offer a definitive recommendation on what to do but rather to illustrate how the concept of data about humans could be operationalized.

4.1 Conceptual Advance: Data About Humans

As discussed above, the personal/non-personal data distinction works well in preventing unwanted disclosure types of data-related harms. Simultaneously, it not only fails to capture the detrimental use types of harms but might, unintendedly, contribute to their emergence. For this reason, to account for the harms to individuals stemming from data-driven knowledge generation and application, policymakers should enrich their conceptual framework with the notion of "data about humans." Its working definition is as follows:

> Data about humans means any information concerning humans, regardless of whether these humans are identified or identifiable, including both data about humans as individuals (including anonymized data) and about humans as members of groups (including aggregated and statistical data).

The relationship of data about humans (and its opposite, i.e., "data not about humans") to personal/non-personal data could be illustrated as illustrated as in Figure 3.

Consequently, all personal data and some non-personal data would be treated as data about humans, whereas some non-personal data would be treated as data not about humans. For example, anonymous data sets containing the shopping history of non-identifiable individuals would fall within the notion of data about humans, as they can be used to detect patterns in human behavior and potentially predict and influence future human behavior, including in a manipulative or discriminatory manner. At the same time, data about performance of machinery, including its longevity and energy consumption, would not be treated as data about humans (unless, from its analysis, one can infer information about human behavior). Similarly, statistical data about the relationship of age, gender, and domicile to political opinions would be data about humans, whereas data about weather would not (unless analyzed or used in some specific contexts, where it can be linked to an identified or identifiable individual).

Several objections to the concept of "data about humans" could be raised. First, the "aboutness" could signify that what matters is the content of information rather than its relation to the natural person(s), and thereby limit the scope of application of personal data protection laws. As it is not proposed to modify the GDPR, the latter consequence would not materialize. However, regarding the former part of the objection, this is precisely the intention of the concept. It aims to be constructed around the content of data, unlike the concept of personal information which focuses on the relationship to the individual. Statistical data about humans, as demonstrated above, can be used in harmful manners and therefore its analysis and usage should be governed by different principles than data about other subjects. Does this mean that data about machinery or weather cannot be used to harm individuals or communities? Of course not. Depending on the context, any data can be used in a harmful manner. However, this does not mean that the concept is not needed. Its introduction is nec-

[114] Kresman, Dunning and Lu (n 73).
[115] Sengupta, Ruj and Bit (n 72).

essary precisely given, among other reasons, the all-encompassing nature of the "personal data."

Second, admittedly, the distinction is not perfectly clear-cut. For example, data about performance of IoT devices, like the longevity of printers or cars, is indirectly linked to human behavior: how often do they use the printers, and how do they drive the cars? Even data about weather could be deemed data about humans, as weather is affected by climate change, in turn driven by the humans emitting CO_2.

However, the usefulness of the concept stems not from its sharp boundaries but its ability to illuminate policymaking. The claim of this article is neither that all data about humans should be treated in the same manner nor that all problems stemming from detrimental use of data about humans can be solved using one regulatory intervention. On the contrary.

4.2 The Concept in Action: Updating Data Law and Policy

When thinking of reforming data law and policy – for example, to address the harms discussed in this article – one faces two questions regarding the regulatory strategy. First, should the interventions be general or issue-specific? Second, should they be technology-centered or area-of-life centered?

The recent legislative action in the EU has opted for the former in both questions. The EU has passed the GDPR (applying to all personal data processing), the Data Governance Act (applicable to all data sharing), and the Digital Services Act[116] (governing all online platforms), debates the Data Act (which would regulate all data generated by the IoT devices) and the Artificial Intelligence Act[117] (imposing obligations on all entities placing AI systems on the EU market). However, this is not the only possible course of action and not always the best one.

First, consider the choice between the general and the issue-specific approach. As the five years of experience with the GDPR illustrate, the horizontal approach can lead to simultaneous under- and overregulation. Of course, the approach has its advantages – no one can escape regulation by clever legal maneuvering – but it also comes with costs. If one works at a university in the EU and wants to start a mailing list or use Zoom for online events, one will have a difficult time. If the university's DPO takes the GDPR seriously, one will need to conclude a data processing agreement with the providers of such services, fulfill all the information obligations, and keep track of all the data processing for accountability purposes. This is costly in money and effort, and arguably unnecessary from the point of view of the individuals who want to sign up for the mailing list or attend an online seminar. At the same time, large corporations like Meta or Google continue to amass and use large amounts of personal data, thanks to large teams of lawyers exploiting holes in the horizontal regulation.[118]

Not all data is equal, and not all uses are equally harmful. Consequently, there is probably no need to adopt anything like a General Regulation on Data About Humans. The available regulatory tools: transparency obligations, individual rights, principles like purpose limitation, need for certification, administrative oversight, etc., might be reasonable in some cases (e.g., personalizing advertisements) while overstepping in others (e.g., monitoring the volume of traffic on one's website to optimize the servers' capacity) and insufficient in yet others ones (like prevention of harms to people's mental health). The approach to data about humans' usage in areas where the potential harm is clear: ad and content personalization, price discrimination, optimizing for engagement, etc., might be very different than in areas where the expected harm is minimal.

Second, consider the choice between technology-centered and area-of-life-centered interventions. One could imagine the legislator adopting something to the effect of a Data-Driven Manipulation Act, a Data-Driven Discrimination Act, etc. However, one could just as well imagine a data-conscious update of consumer law, non-discrimination law, employment law, etc. The latter approach is prudent when the regulatory goals are not immediately clear. Within the existing logic, founded on the personal/non-personal distinction, there is a risk of, e.g., updating the consumer law to be more mindful of personal data processing. One could assume that, as long as non-personal data about humans (e.g., statistical data) is processed, the risk of harm is absent. The concept of "data about humans" aims to eliminate this risk.

Simply identifying a problem – for example, enumerating the harms like in this article – does not automatically translate into a clear-cut set of policy goals. Think back to the harms discussed in the previous section: discrimination in ad-delivery,[119] price discrimination,[120] manipulation[121] or negative impacts on mental health.[122] From the point of view of the individual affected, these are always problematic. What to do about them, however, is a political choice.

Should we ban online advertising, limit the groups who are allowed to receive them, or limit the kinds of techniques advertisers are allowed to employ? Those are difficult questions, and the expert knowledge needed to ponder the pros and cons and various approaches is not (only) the understanding of data science but also of consumer law. Consumer law pursues several goals at the same time[123] and, in the face of a new sociotechnological reality, might want to reconsider the balancing of these goals adopted in the past. The choice is not a choice about technology but about the shape of the society we want to live in.

Similarly, hard choices are present in all areas of harms. If we want to fight discrimination in ad delivery, what would be the acceptable end goal? Do we want the groups receiving ads to mirror the demographics of the society, or do we want to ensure some minimal quotas? To answer such questions, one needs to go deep into the philosophy of antidiscrimination laws.[124] Regarding price discrimination: do we want to ban it altogether? Or do we want to ban certain excesses lowering the efficiency of the market? Here we need insights from the economic law regarding the function of

[116] 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), *OJ L 277, 27.10.2022, p. 1–102.*

[117] The text is changing, official updates available at https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.

[118] Matt Burgess, 'How GDPR Is Failing' [2022] Wired <https://www.wired.com/story/gdpr-2022/> accessed 26 August 2022.

[119] Sweeney (n 9); Prince and Schwarcz (n 16); Ali and others (n 9); Xenidis (n 9).

[120] Woodcock (n 10); Bar-Gill (n 10).

[121] Mik (n 8); Susser, Roessler and Nissenbaum (n 8); Trzaskowski (n 105).

[122] Wu (n 86); Hou and others (n 99); Zakon (n 11).

[123] Herrine (n 55); Herrine (n 58); Helberger and others (n 106).

[124] Xenidis (n 9).

prices.[125] Hence, instead of regulating data usage in various areas of life, we might want to update the existing laws already governing those areas. Such an update needs to be undertaken with a profound understanding of how data analytics works and the role that data about humans plays there. Choices regarding the regulatory approaches, the policy goals, and the suitable means will be different depending on the potential for harm, the kind of social good in danger, and the potential cost of intervention.

The concept of data about humans is useful not because it is clear-cut. Its power lies in, first, deconstructing the false conviction that individuals can only be harmed when processing personal data and, second, bringing the role of non-personal data analytics to the forefront. It is not a concept that ever needs to be enshrined in the positive law, even though it might be useful to do so in some cases. However, it should make its way to policymakers' toolbox used for analyzing the drawbacks of digitalization and the possible interventions for addressing them.

5. Conclusion

The goal of this article has been to demonstrate how the personal/non-personal distinction in data law and policy makes it difficult to appreciate the extent to which individuals and social groups might be harmed by non-personal data processing. This distinction currently structures data laws worldwide and gives rise to a dangerous mental shortcut: personal data processing comes with potential harms to individuals and should be constrained, whereas non-personal data processing leads only to generation of new value and should be encouraged. It has been demonstrated that, when it comes to detrimental use types of harms, non-personal data processing can lead to discrimination in access, price discrimination, manipulation, and negative impacts on mental health.

Consequently, this article suggests the adoption of a new concept, namely "data about humans." This category would encompass all data about human behavior, regardless of whether it concerns identifiable or anonymous persons, and include both data about individuals and groups. The concept is an intellectual tool. It does not have to make its way to any binding law, though there might be situations when this is beneficial. Rather, it is supposed to enrich the toolbox for scholars and policymakers pondering the potential harms resulting from the emergence of the data society and economy, as well as the means for addressing them.

Indeed, it is possible to structure data laws and policies in a way that taps into the potential of digitalization by incentivizing value creation while protecting individuals from harms. However, the distinction between personal and non-personal data is ill-suited to inform the regulatory strategy. It should be complemented by the notion of data about humans.

[125] Woodcock (n 10); Bar-Gill (n 10).