The Legal Protection Debt of Training Datasets

Gianmarco Gori
Postdoctoral Researcher at the Law, Science, Technology and Society Research Group (LSTS),
Vrije Universiteit Brussel (VUB)

HumanE AI Net
European Union's Horizon 2020 research and innovation programme
Grant Agreement No 952026

Executive summary

LSTS (VUB) is a partner of the HumanE-AI Project, participating to WP5 on the legal and ethical bases for responsible AI. Under T5.2., LSTS visits microprojects and interacts with AI researchers to engage in a constructive technology assessment, teasing out potential risks for the rights and freedoms of natural persons who may suffer the consequences of implementation of AI.

The present Report is the result of the involvement of LSTS in the microproject "Collection of datasets tailored for HumanE-AI multimodal perception and modelling". Such microproject has been carried out in the context of WP2, Multimodal Perception and Modeling. The goal of WP2 is to provide integrated multi-modal perception and modelling to develop systems that can understand complex human actions, motivations and social settings. The microproject aimed at contributing to the research field by putting at disposal of the scientific community ready-to-use, curated datasets for multimodal perception and modelling of human activities and gestures. The partners involved have created, curated and released datasets for Human Activity Recognition (HAR) tasks, in particular, the extended dataset OPPORTUNITY++ and Wearlab BeachVolleyball dataset.

The participation to the microproject has offered the chance to get a closer look at the practices, doubts and difficulties emerging within the scientific community involved in the creation, curation and dissemination of training datasets. Considering that one of the goals of the HumanE-AI Net is to connect research with relevant use cases in European society and industry, the participation to the microproject has offered the occasion to situate dataset collection, curation, and release within the broader context of AI pipeline.

Under T.5.2., the task of LSTS is to provide researchers with recommendations on how to integrate legal protection by design into the architectures developed in the microprojects. To this end, the Report examines the potential issues that arise within current ML-practices and provide an analysis of the relevant normative frameworks that govern such practices. By bridging the gap between practices and legal norms, the Report provides researchers with the tools to assess the risks to fundamental rights and freedoms that may occur due to the implementation of AI research in real world situations and recommends a set of mitigating measures to reduce infringements and to prevent violations.

The Report acknowledges that datasets constitute the backbone infrastructure underpinning the development of Machine Learning. The datasets that are created, curated and disseminated by ML practitioners provide the data to train ML models and the benchmarks to test the improvement of such models in performing the tasks for which they are intended. However, until recently, the practices, processes and interactions that take place further upstream the ML-pipeline, between the collection of data and the use of dataset for training ML-models, have tended to fade into the background.

The report argues that the practices of dataset creation, curation and dissemination play a crucial role in the setting of the level of legal protection that is afforded to all the legal subjects that are located downstream ML-pipelines. Where such practices lack appropriate legal safeguards, a "Legal Protection Debt" can mount up incrementally along the stages of ML-pipelines. In section 1.1., the Report provides a brief overview of how current data science practices depend on and perpetuate an ecosystem characterised by a lack of structural safeguards for the risks posed by data processing. This can lead to the accumulation of "technical debt". Such debt, in turn, can assume relevance in the perspective of the compliance with legal requirements. Taking inspiration from the literature on technical and ethical debt, the Report introduces the concept of Legal Protection Debt. Because of this legal protection debt, data-driven systems implemented at the end of the ML pipeline may lack the safeguards necessary to avoid downstream harm to natural persons.

The Report argues that the coming about of Legal Protection Debt and its accumulation at the end of the ML pipeline can be addressed through the adoption of a Legal protection by design approach. This implies the overcoming of a siloed understanding of legal liability that mirrors the modular character of ML pipelines. Addressing legal protection debt requires ML practitioners to adopt a forward looking perspective. Such perspective should situates the stage of development in practitioners are involved in the context of the further stages that take place both upstream and downstream the pipeline. The consideration of the downstream stages of the ML-pipeline shall, as it were, backpropagate and inform the choices as to the technical and organisational measure to be taken upstream: upstream design decisions must be based on the anticipation of the downstream uses afforded by datasets and the potential harms that the latter may cause. Translated into a legal perspective, this implies that the actors upstream the pipeline should take into consideration the legal requirements that apply to the last stages of the pipeline.

The Report illustrates how data protection law lays down a set of legal requirements that overcome modularity and encompass the ML pipeline in its entirety, connecting the actors upstream with those downstream. The GDPR makes controllers responsible for the effects of the processing that they carry out. In section 2, the Report shows how the GDPR provides the tools to mitigate the problem of many hands in ML-pipelines. The duties and obligations set by the GDPR require controllers to implement by design safeguards that conjugate the need to address downstream harms with the necessity to comply with the standards that govern scientific research. In this perspective, the Report shows that the obligations established by data protection law either instantiate or

harden most of the requirements set by the Open science and Open data framework and also the best practices emerging within the ML-community.
In section 2.1. the Report illustrates the core structure of the regime of liability to which controllers are subject under the GDPR. Such regime of liability hinges upon controllers' duty to perform a context-dependent judgment. Such judgment must informs controllers' decisions as to the measures to be adopted to ensure compliance with all the obligations established by the GDPR. Such judgment must be based on the consideration of the downstream harms posed by the processing.
In essence, the duty to anticipate and address potential downstream harms requires controllers to adopt a forward-looking approach. In order to ensure compliance with the GDPR, controllers must engage in a dynamic, recursive practice that addresses the requirements of present processing in the light of the future potential developments. At the same time, the planning effort required by the GDPR is strictly connected with the compliance with obligations set by other normative frameworks. In this sense, compliance with the GDPR and compliance with obligations such as those imposed by the Open science and Open data framework go hand in hand. Compliance with the GDPR is a pre-requisite for complying with Open science and Open data framework. Simultaneously, the perspective of open access and re-usability of datasets affects the content of the obligations set by the GDPR.
As a result, the consideration of "what happens downstream" - i.e., the potential uses of datasets, potential harms that the latter may cause, further requirements imposed by other normative frameworks – backpropagates, determining the requirements that apply upstream.
In section 2.2. we show how the compliance with the documentation obligations set by the GDPR can address the accumulation of a documentation debt and ensure controllers' compliance with the obligations established by other normative framework, such as Open Data and Open Science. The overlapping between the documentation requirements established by such different frameworks shows firstly that a serious approach to the compliance with the GDPR can provide the safeguards necessary to avoid the accumulation of a documentation debt. In this way, compliance with the documentation obligations set by the GDPR can prevent the accumulation of other forms of technical debt and, eventually, of legal protection debt. At the same time, the convergence between the requirements set by the GDPR and those established by the FAIR principle and the Horizon DMP template shows how the performance of the documentation obligations established by the GDPR can also facilitate compliance with requirements specific to data processing conducted in the context of scientific research.
A correct framing of the practices of dataset creation, curation and release in the context of research requires to make an effort towards the integrity of the legal framework as a whole, taking into consideration the relations between Open data, Open science and data protection law. First, it is first important to stress that compliance with data protection law represents a pre-requisite for the achievement of the goals of Open Data and Open Science framework.
In section 2.3. the Report analyses the requirements that govern the release and downstream (re)use of datasets. Compliance with the requirements set by the GDPR is essential to avoid that dataset dissemination gives rise to the accumulation of legal protection debt along ML pipelines. Based on the assessment of adequacy and effectiveness required for all forms of processing, controllers can consider the adoption a range of measures to ensure that data transfer are compliant with the GDPR. Among such measures, the Report examines the use of licenses, the providing of adequate documentation for the released dataset, data access management and traceability measures, included the use of unique identifiers.
The Report contains an Annex illustrating the provisions of the GDPR that establish a special regime for the processing carried out for scientific research purposes. We highlight how most of the provisions contained in the GDPR are not subject to any derogation or exemption in view of the scientific research purpose of the processing. All in all, the research regime provided by the GDPR covers the application of a limited number of provisions (or part of provisions). A processing that is unlawful in that it does not comply with the general provisions set by the GDPR cannot enjoy the effects of the derogations provided by the research regime. The derogations allowed under the special research regime concern almost exclusively the GDPR provisions on the rights of data subjects, while no derogation is possible for the general obligations that delineate the responsibility of the controller. The derogations provided under the special research regime allow controllers to modulate their obligations towards data subjects where the processing of personal data is not likely to affect significantly the natural persons that are identified or identifiable through such data. As it were, the decrease of the level of potential harm makes possible the lessening of the safeguards required to ensure the protection of data subjects. Even in such cases, however, no derogation is allowed with respect to the requirements different than those concerning the rights of data subject. This circumstance makes manifest that the system established by the GDPR aims at providing a form of protection that goes beyond the natural persons whose personal data are processed at that time by controllers.

Table of Contents

Introduction

LSTS (VUB)[1] is a partner of the HumanE-AI Project[2], participating to WP5 on the legal and ethical bases for responsible AI. Under T5.2., LSTS visits microprojects and interacts with AI researchers to engage in a constructive technology assessment, teasing out potential risks for the rights and freedoms of natural persons who may suffer the consequences of implementation of AI.

The present Report is the result of the involvement of LSTS in the microproject "Collection of datasets tailored for HumanE-AI multimodal perception and modelling" [3]. Such microproject has been carried out in the context of WP2, Multimodal Perception and Modelling. The goal of WP2 is to provide integrated multi-modal perception and modelling to develop systems that can understand complex human actions, motivations and social settings[4]. The microproject aimed at contributing to the research field by putting at disposal of the scientific community ready-to-use, curated datasets for multimodal perception and modelling of human activities and gestures[5]. The partners involved have created, curated and released datasets for Human Activity Recognition (HAR) tasks[6], in particular, the extended dataset OPPORTUNITY++[7] and Wearlab BeachVolleyball dataset[8].

The field of Human Activity Recognition is based on the collection of datasets composed of sensor, video and synthetic data[9]. Such datasets are used to train ML-algorithms that aim at modelling the human behaviour represented in data. The ML-models, in turn, are used to build systems that aim at recognising human behaviour in specific activities and situations. HAR has multiple applications in contexts that have been extensively under the attention of legal scholars and institutions especially in the light of data protection law[10], such as Ambient Intelligence and the Internet of Things (IoT). HAR

---

[1] https://lsts.research.vub.be/

[2] https://www.humane-ai.eu/

[3] https://www.humane-ai.eu/project/collection-of-datasets-tailored-for-humane-ai-multimodal-perception-and-modelling/; D1.1: First Year Microproject Results from Workpackage 1, 2 and 3, pp. 26-27, https://www.humane-ai.eu/wp-content/uploads/2021/10/Deliverable-1.1.pdf

[4] https://www.humane-ai.eu/workpackages/

[5] D1.1: First Year Microproject Results from Workpackage 1, 2 and 3, pp. 26-27, https://www.humane-ai.eu/wp-content/uploads/2021/10/Deliverable-1.1.pdf

[6] Ivi

[7] Mathias Ciliberto and others, 'Opportunity++: A Multimodal Dataset for Video- and Wearable, Object and Ambient Sensors-Based Human Activity Recognition' (2021) 3 Frontiers in Computer Science <https://www.frontiersin.org/articles/10.3389/fcomp.2021.792065> accessed 28 April 2023. Daniel Roggen and others, 'Collecting Complex Activity Datasets in Highly Rich Networked Sensor Environments', *2010 Seventh International Conference on Networked Sensing Systems (INSS)* (2010). https://doi.org/10.1109/INSS.2010.5573462. A list of publications relative to the Opportunity dataset is available at http://www.opportunity-project.eu/node/56.html

[8] Mathias Ciliberto, Luis Alejandro Ponce Cuspinera and Daniel Roggen, 'Collecting a Dataset of Gestures for Skill Assessment in the Field: A Beach Volleyball Serves Case Study', *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (Association for Computing Machinery 2021) <https://doi.org/10.1145/3460418.3479355> accessed 28 April 2023.
Daniel Roggen, Arash Pouryazdan and Mathias Ciliberto, 'Poster: BlueSense - Designing an Extensible Platform for Wearable Motion Sensing, Sensor Research and IoT Applications'. In Proceedings of the 2018 International Conference on Embedded Wireless Systems and Networks. ACM.

[9] Kaixuan Chen and others, 'Deep Learning for Sensor-Based Human Activity Recognition: Overview, Challenges, and Opportunities' (2021) 54 ACM Computing Surveys 77:1. https://doi.org/10.1145/3447744; Santosh Kumar Yadav and others, 'A Review of Multimodal Human Activity Recognition with Special Emphasis on Classification, Applications, Challenges and Future Directions' (2021) 223 Knowledge-Based Systems 106970. https://doi.org/10.1016/j.knosys.2021.106970; Preksha Pareek and Ankit Thakkar, 'A Survey on Video-Based Human Action Recognition: Recent Updates, Datasets, Challenges, and Applications' (2021) 54 Artificial Intelligence Review 2259. https://doi.org/10.1007/s10462-020-09904-8; L Minh Dang and others, 'Sensor-Based and Vision-Based Human Activity Recognition: A Comprehensive Survey' (2020) 108 Pattern Recognition 107561.doi.org/10.1016/j.patcog.2020.107561; Florenc Demrozi and others, 'Human Activity Recognition Using Inertial, Physiological and Environmental Sensors: A Comprehensive Survey' (2020) 8 IEEE Access 210816. https://doi.org/10.1109/ACCESS.2020.3037715

[10] Ex multis: Article 29 Working Party, Opinion 8/2014 on the on Recent Developments on the Internet of Things, 14/EN WP 223, adopted on 16 September 2014, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp223_en.pdf; Mauritius Declaration on the Internet of Things, 36th International Conference of Data Protection and Privacy Commissioners, Balaclava, 14 October 2014, https://edps.europa.eu/sites/default/files/publication/14-10-14_mauritius_declaration_en.pdf; Christos Karageorgiou Kaneen and Euripides GM Petrakis, 'Towards Evaluating GDPR Compliance in IoT Applications' (2020) 176 Procedia Computer Science 2989. https://doi.org/10.1016/j.procs.2020.09.204; Nóra Ni Loideain, 'A Port in the Data-Sharing Storm: The GDPR and the Internet of Things' (2019) 4 Journal of Cyber Policy 178. DOI: 10.1080/23738871.2019.1635176; Sandra Wachter, 'The GDPR and the Internet of Things: A Three-Step Transparency Model' (2018) 10 Law, Innovation and Technology 266. DOI: 10.1080/17579961.2018.1527479

underpins all the forms of computation that build on computer-human interactions based on human gestures, with use cases spanning from domotics, to skill assessment, care services, healthcare, industry, sport, etc.[11].

By its very name, Human Activity Recognition involves forms of data processing that are inherently based upon and directed to natural persons. Personal data are present within the entire HAR pipeline: from the dataset collection and curation to the deployment of real-world applications, sensor, video, etc. data are about natural persons. At the same time, the processing of personal data is not centred on the identification or identifiability of the natural persons from which data are collected. While HAR relies upon personal data, the data produced by the single data subjects assume relevance as instances of generalisable segments of behaviour, resulting in a weak link between data and subject. Datasets composed of personal data are used to train ML models. The latter are likely to be implemented into systems that are not going to affect the data subjects whose data have been collected into the original dataset. Other data subjects will be affected by the technologies implementing ML models that have been trained on the dataset. Yet, the data subjects whose data are collected into datasets are, at least, identifiable, thereby triggering the application of data protection law. At the same time, the activities performed in the context of the microproject assume relevance under other normative frameworks, such as Open data and Open science. The involvement in the microproject has provided the opportunity to examine the intersection between the challenges posed by dataset creation, curation and dissemination and the multiple normative frameworks that govern such practices. The analysis conducted has highlighted that many of the features that characterise the field of HAR are common to multiple area of ML research. Considering that one of the goals of the HumanE-AI Net is to connect research with relevant use cases in European society and industry[12], the participation to the microproject has offered the occasion to situate dataset related practices within the broader context of the AI pipeline. Under T.5.2., the task of LSTS is to provide researchers with recommendations on how to integrate legal protection by design into the architectures developed in the microprojects. To this end, the Report examines the potential issues that arise within current ML-practices and provide an analysis of the relevant normative frameworks that govern such practices. By bridging the gap between practices and legal norms, the Report aims at providing researchers with the tools to assess the risks to fundamental rights and freedoms that may occur due to the implementation of AI research in real world situations.

The report argues that the practices of dataset creation, curation and dissemination play a crucial role in the setting of the level of legal protection that is afforded to all the legal subjects that are located downstream ML-pipelines. Where such practices lack appropriate legal safeguards, a “Legal Protection Debt” can mount up incrementally along the stages of ML-pipelines.

The finding of the research highlight the need that the actors involved in the AI pipeline adopt a forward-looking perspective to legal compliance. This requires the overcoming a siloed and modulated understanding of legal liability and the paying of keen attention to the potential use cases of datasets produced in research contexts.

The Report is structured as follows.

Section 1 provides a brief overview of the structural features of current practices of dataset creation, curation and dissemination. Section 1.1. illustrates how the current data(set) ecosystem is characterised by a lack of structural safeguards addressing the risks that arise along ML pipelines. The Report illustrates how the incentive structure that characterise the data ecosystem can result in the accumulation of “technical debt”. Such debt, in turn, can produce a Legal Protection Debt, that is, a debt at the level of legal protection enjoyed by natural persons downstream the ML pipeline. Section 1.2. outlines of how the collection, curation and release of datasets can be addressed under the framework established by the GDPR.

Section 2 illustrates how data protection law lays down a set of legal requirements that can help overcoming the challenges examined in Section 1. Section 2.1. shows how the duties and obligations set by the GDPR require controllers to implement by design safeguards that conjugate the need to address downstream harms with the necessity to comply with the standards that govern scientific research. Section 2.2. examines how the compliance with the documentation obligations set by the GDPR can address the accumulation of a documentation debt and ensure controllers’ compliance

[11] Cf, nn 7-9

[12] WP 6: Applied research with industrial and societal use cases, https://www.humane-ai.eu/workpackages/

with the obligations established by other normative framework, such as Open Data and Open Science. Section 2.3. analyses the requirements that govern the release and downstream (re)use of datasets. Compliance with the requirements set by the GDPR is essential to avoid that dataset dissemination gives rise to the accumulation of legal protection debt along ML pipelines.
The Conclusions summarise the recommendations made in the Report, indicating a set of measures aimed at implementing legal protection by design in ML pipelines.
The Annex provides an analysis of the special regime established by the GDPR for the cases in which the processing of personal data is performed for scientific research purposes.

1. Legal Protection Debt in dataset practices

Datasets constitute the backbone infrastructure underpinning the development of Machine Learning[13]. The dataset that are created, curated and disseminated by ML practitioners provide the data to train ML models and the benchmarks to test the improvement of such models in performing the tasks for which they are intended[14]. Yet, until recently, most of the attention paid by both policy makers and academics to ML-pipelines has been polarised to the extremities of the latter: upstream, much attention has been paid to the practices of data collection; downstream, the public and scientific debate has addressed the stage in which ML model are implemented into systems that produce effects on legal subjects. Upstream, the moment in which data are collected has received attention especially in the light of data protection and privacy law, raising concerns focused on the requirement of consent or other legal basis of data processing. Downstream, the harms produced by data-driven systems have among others, discrimination, and explainability and justifiability of automated decision making. In this sense, the attention to the effects of data-driven system has stimulated the interest into the upstream stages of the ML-pipeline, namely the stage in which datasets are used to train ML-models. However, until recently, the practices, processes and interactions that take place further upstream, between the collection of data and the use of dataset for training ML-models have tended to fade into the background.
In this section we argue that the practices of dataset creation, curation and dissemination play a crucial role in the setting of the level of legal protection that is afforded to all the legal subjects that are located downstream ML-pipelines. Where such practices lack appropriate legal safeguards, a “Legal Protection Debt” can mount up incrementally along the stages of ML-pipelines. We claim that a dedicated analysis the current legal framework illustrates how the law in force establishes a set of requirements that apply to dataset practices seamlessly along the entire ML-pipeline. We claim that compliance with the obligations set by the current legal framework requires ML practitioners to adopt measures that address the accumulation of Legal Protection Debt.
In section 1.1., we provide a brief overview of how current data science practices depend on and perpetuate an ecosystem characterised by a lack of structural safeguards for the risks posed by data processing. We then discuss how the challenges posed by dataset usage can be framed under a legal perspective and, in particular, under data protection law. In section 1.2. we provide an outline of how the collection, curation and release of datasets can be addressed under the framework established by the GDPR.

1.1. Datasets practices and the ML-pipeline: modularity, “many hands”, and debts.

The practices of dataset creation, curation and release have started attracting growing attention in connection to recent retraction of popular datasets in the field of computer vision and natural language

[13] Ben Hutchinson and others, ‘Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure’, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2021) <https://dl.acm.org/doi/10.1145/3442188.3445918> accessed 28 April 2023, at 560.
[14] Inioluwa Deborah Raji and others, ‘AI and the Everything in the Whole Wide World Benchmark’ (arXiv, 26 November 2021) <http://arxiv.org/abs/2111.15366> accessed 28 April 2023.The Authors define benchmark as “a particular combination of a dataset or sets of datasets (at least test data, sometimes also training data), and a metric, conceptualized as representing one or more specific tasks or sets of abilities, picked up by a community of researchers as a shared framework for the comparison of methods”.

processing (NLP)[15]. In the recent years, an increasing number of studies have addressed the life cycle of datasets, from their development to their (re)use[16]. Such studies illustrate how, both in scientific research and industry, the current ML practice is rests upon the open availability and re-usability of software, models, datasets[17]. This has favoured the growing diffusion of parent and derivative datasets, making the number of dataset published in online repository hard to estimate[18]. At the same time, dataset usage tends to assume a long tail distribution within ML communities, due to the convergence of practitioners' work on a small group of datasets[19]. While dataset usage is concentrated on a small group of datasets, the complexity of the dataset lifecycles is increased by the circumstance that successful datasets tend to be re-used for tasks different from those for which they were created[20]. The current practice reflects values characteristic of scientific research. In this sense, the EU policy endorses the idea of data science and innovation as a collaborative effort based on progress build upon the research and development efforts made by others[21]. As the European Commission affirms in the European Data Strategy, "the value of data lies in its use and re-use. Currently there is not enough data available for innovative re-use, including for the development of artificial intelligence. …Several of the issues concern the availability of data for the public good"[22]. In a similar vein, the Open data directive highlights that "open acces to data helps enhance quality, reduce the need for unnecessary duplication of research, speed up scientific progress, combat scientific fraud, and it can overall favour economic growth and innovation"[23].

Whereas most of the abovementioned studies focus on the field of computer vision and NLP, the features of dataset usage thereby identified seem to characterise ML practice irrespective of the specific field of research, including Human Activity Recognition[24]. A common structure of ML-pipelines seem to prevail across different sectors of data science. The work of practitioners involved in dataset creation, curation, release and (re)use is divided according to a modular structure. In its multiple modules, the ML-pipeline can involve several actors. Such actors can belong to different sub-communities within data science and be distributed across the research and industrial sector. The

[15] Such as Tiny Images, VGGFace2, DukeMTMC, and MS-Celeb-1M. See the literature below.

[16] Emily Denton and others, 'On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet' (2021) 8 Big Data & Society 20539517211035956. 1–14. DOI: 10.1177/20539517211035955; Jesse Dodge and others, 'Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus' (arXiv, 30 September 2021) <http://arxiv.org/abs/2104.08758> accessed 28 April 2023; Kate Crawford and Trevor Paglen. 2019. Excavating AI: The politics of training sets for machine learning. https://www.excavating.ai; Amandalynne Paullada and others, 'Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research' (2021) 2 Patterns 100336.

[17] Suzanne L Thomas, 'Migration Versus Management: The Global Distribution of Computer Vision Engineering Work', *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)* (2019). https://doi.org/10.1109/ICGSE.2019.00017.

[18] Omar Benjelloun, Shiyu Chen, and Natasha Noy, Google Dataset Search by the Numbers, Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, Lalana Kagal (Eds.), The Semantic Web – ISWC 2020. 19th International Semantic Web Conference Athens, Greece, November 2–6, 2020 Proceedings, Part II, Springer, 2020, pp 667–682. https://doi.org/10.1007/978-3-030-62466-8

[19] Bernard Koch, Emily Denton, Alex Hanna, Jacob G Foster, Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research, Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021)

[20] Ivi

[21] In this sense, e.g., Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A European strategy for data, COM(2020) 66 final, Brussels, 19.2.2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066; Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information, C/2018/2375, http://data.europa.eu/eli/reco/2018/790/oj; Council conclusions on open, data-intensive and networked research as a driver for faster and wider innovation, 9360/15, Brussels, 29 May 2015, https://data.consilium.europa.eu/doc/document/ST-9360-2015-INIT/en/pdf; Council conclusions on the transition towards an open science system, 9526/16, Brussels, 27 May 2016, https://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf

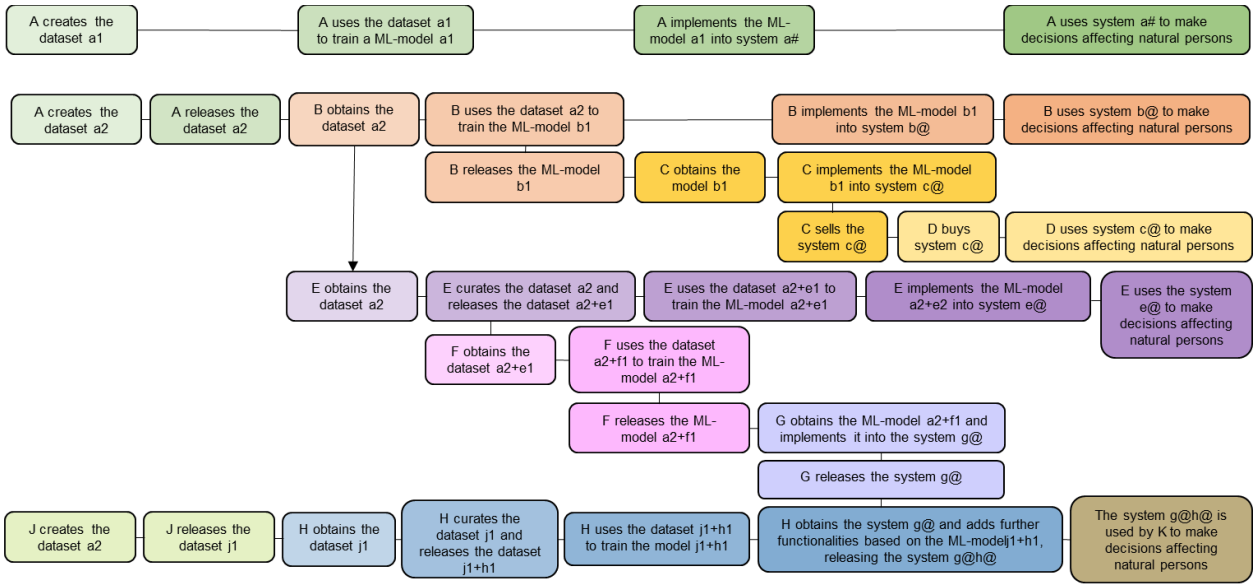
[22] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A European strategy for data, COM(2020) 66 final, Brussels, 19.2.2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066, p. 6

[23] 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), http://data.europa.eu/eli/dir/2019/1024/oj (hereafter, Open data directive, or ODD), Recital 2; Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information, C/2018/2375, http://data.europa.eu/eli/reco/2018/790/oj, § 9: "Member States should ensure that, as a result of those policies or action plans…the academic career system supports and rewards researchers who participate in a culture of haring the results of their research, in particular by ensuring early sharing and open access to their publications and other research outputs".

[24] Minh Dang and others (n 9); Yadav and others (n 9); Pareek and Thakkar (n 9).

multiple actors involved in the pipeline can be potentially distant from each other, depending on where and how much the modules in which they work are situated upstream or downstream the pipeline[25].
The modular and distributed structure of the ML-pipeline mirrors and partly overlaps with the supply chain characteristic of software engineering. The similarities between the two fields is not limited to the structure of the work-chain. ML-practice shares with software engineering also the values that correspond to such structure[26]. As for the engineering perspective, modularity is promoted for its advantages in terms of efficiency, specialisation of work, scalability, incremental re-usability.
While affording several benefits, the current structure of ML-pipelines give rise to several concerns. In this perspective, the parallel between ML practices and software engineering applies also to the other side of the coin. The structure of ML-pipelines reproposes and exacerbates the challenges emerged with respect to "classic" software engineering supply chains[27]. As represented in fig. 1, "many hands" can operate between the modules upstream and the downstream deployment of data-driven systems. Modular and distributed pipelines can be affected by the problem of many hands[28] and favour the handoff of responsibility from modules upstream to modules downstream.

Figure 1: "Many hands" in the ML pipeline



In particular, such structure depends on and normalises a certain perception of the responsibility within the practices of dataset creation, curation and release. In this perspective, it is telling that, while accountability and responsibility have been extensively addressed with reference to the training and implementation of ML models[29], only recently the attention has ascended the pipeline towards the practices of the actors upstream involved in dataset creation, curation and release. In the last years, qualitative studies have highlighted how practitioners involved in modular and distributed ML-pipelines tend to develop a siloed understanding of responsibility. Such a siloed understanding of responsibility reflects the rigid division of the work between modules and increases the more the module in which the practitioner works is upstream the pipeline. The closer practitioners are to the final deployment of

[25] David Gray Widder and Dawn Nafus, 'Dislocated Accountabilities in the AI Supply Chain: Modularity and Developers' Notions of Responsibility' (arXiv, 27 September 2022) <http://arxiv.org/abs/2209.09780> accessed 28 April 2023.
[26] Hutchinson and others (n 14). Widder and Nafus (n 26).
[27] A Feder Cooper and others, 'Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning', *2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) <https://doi.org/10.1145/3531146.3533150> accessed 28 April 2023. Helen Nissenbaum. 1996. Accountability in a computerized society. Sci Eng Ethics 2, 1 (March 1996), 25–42. https://doi.org/10.1007/BF02639315
[28] The first formulation of the "problem of many hands" was proposed by Dennis F. Thompson to describe the difficulty to identify the agent that is moral responsible for political outcome in the circumstances in "which many different officials contribute in many ways to decisions and policies of government". Dennis F Thompson, 'Moral Responsibility of Public Officials: The Problem of Many Hands' (1980) 74 American Political Science Review 905.DOI: https://doi.org/10.2307/1954312.
[29] Cooper and others (n 28).

a system, the more they tend to be sensitive to potential harmful impact of their work[30]. The more upstream, the more "[t]he celebration of endless possibilities for downstream use makes harms appear to be a general, unconnected matter"[31]. Widder and Nafus have found that the practitioners working more upstream the pipeline tend to account for their action by reference to "discourses of technological neutrality": practitioners refer to "what they make as not even comparable to the gun that proverbially can used for both harm or good, but the machinery which makes multi-use parts: 'I make a piece of equipment that makes pipe, somebody bought my pipe making equipment, and made the barrel of guns. I don't know how I stop [harm], because I didn't make the gun'"[32]. The perception of a lack of control on, and responsibility for, the downstream lifecycle of datasets is amplified by the perspective of the potentially unrestricted circulation of datasets favoured under the Open source and Open data framework[33].

This perception of (ir)responsibility is self-fulfilling, in that it favours the adoption of irresponsible data practices. An approach of the kind "there is nothing that I can do at this stage" can easily lead to a scenario in which, before downstream harms, "nothing can't be done anymore". Modular, distributed and open pipelines risk normalising an *incentive structure* that, by restricting the locus of responsibility to the last stages downstream, results in a lack of effective protection and in an inefficient allocation of resources. Such incentive structures hinders the building of safeguards against data cascades, i.e., "compounding events causing negative, downstream effects from data issues"[34]. A siloed understanding of responsibility and the practices of responsibility handoff can result in the accumulation of multiple "debts" along ML-pipelines.

Such "debts" can consist first of all in "technical debt"[35]. In this sense, the recent literature has emphasised how current dataset practices can give rise to documentation debt[36]. Technical debt, in its various forms, can assume relevance in the perspective of the compliance with legal requirements. The acts or omissions that give rise to technical debt might constitute in themselves a violation of a legal obligations. For instance, documentation debt can result from an actor's failure to form or keep technical documentation specifically prescribed by the law. Legally relevant technical debt can also result from the combined effect of multiple acts that, if taken separately, do not configure a violation of legal requirements, at least prima facie. These cases are more complex and risky, in that the

[30] Widder and Nafus (n 26). Will Orr and Jenny L Davis, 'Attributions of Ethical Responsibility by Artificial Intelligence Practitioners' (2020) 23 Information, Communication & Society 719. DOI: 10.1080/1369118X.2020.1713842; Thomas (n 18).

[31] Widder and Nafus (n 26).

[32] ibid., references omitted

[33] Thomas (n 18). Widder and Nafus (n 26).

[34] Nithya Sambasivan and others, '"Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI', *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2021) <https://doi.org/10.1145/3411764.3445518> accessed 28 April 2023.

[35] ibid. The metaphor of technical debt was introduced by Ward Cunningham in 1992. Ward Cunningham, 'The WyCash Portfolio Management System', *Addendum to the proceedings on Object-oriented programming systems, languages, and applications (Addendum)* (Association for Computing Machinery 1992) <https://dl.acm.org/doi/10.1145/157709.157715> accessed 28 April 2023. See, also, Girish Suryanarayana, Ganesh Samarthyam, Tushar Sharma, Refactoring for Software Design smells. Managing Technical Debt, Elsevier, 2015.

[36] Emily M Bender and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) <https://dl.acm.org/doi/10.1145/3442188.3445922> accessed 28 April 2023, at 615; Mahima Pushkarna, Andrew Zaldivar and Oddur Kjartansson, 'Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI', *2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) <https://dl.acm.org/doi/10.1145/3531146.3533231> accessed 28 April 2023. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford, Datasheets for datasets, Communications of the ACM, December 2021, Vol. 64 No. 12, Pages 86-92, DOI:10.1145/3458723; Hutchinson and others (n 14); Sarah Holland and others, 'The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards' (arXiv, 9 May 2018) <http://arxiv.org/abs/1805.03677> accessed 28 April 2023. In parallel, proposals for dataset documentation have been made with respect to specific sectors, e.g., for datasets in natural language processing: Emily M Bender and Batya Friedman, 'Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science' (2018) 6 Transactions of the Association for Computational Linguistics 587. http://dx.doi.org/10.1162/tacl_a_00041; for datasets in computer vision: Milagros Miceli and others, 'Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2021) <https://dl.acm.org/doi/10.1145/3442188.3445880> accessed 28 April 2023.; for dataset in social computing: R Stuart Geiger and others, 'Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2020) <https://dl.acm.org/doi/10.1145/3351095.3372862> accessed 28 April 2023.

accumulation of the debt can be less visible. As a result, the debt can become evident only at stage in which no ex post measure can provide adequate remedies. Even when the amount of debt accumulated in each stage of a pipeline is marginal, the combination of such debts can in the long run affect the level of legal protection enjoyed by natural persons downstream. Taking inspiration from the literature on technical and ethical debt[37], we propose to refer to this debt as legal protection debt. Because of this legal protection debt, data-driven systems implemented at the end of the ML pipeline may lack the safeguards necessary to avoid downstream harm to natural persons. In the absence of adequate safeguards throughout the entire life cycle of datasets, legal protection debt can become visible – and addressable – only too late.

The coming about of legal protection debt and its accumulation at the end of the ML pipeline can be avoided through the adoption of a Legal protection by design approach[38]. This implies the overcoming of a siloed understanding of legal liability that mirrors the modular character of ML pipelines. Addressing legal protection debt requires ML practitioners to adopt a forward looking perspective. Such perspective should situates the stage of development in practitioners are involved in the context of the further stages that take place both upstream and downstream the pipeline. The consideration of the downstream stages of the ML-pipeline shall, as it were, backpropagate and inform the choices as to the technical and organisational measure to be taken upstream: upstream design decisions must be based on the anticipation of the downstream uses afforded by datasets and the potential harms that the latter may cause. Translated into a legal perspective, this implies that the actors upstream the pipeline should take into consideration the legal requirements that apply to the last stages of the pipeline.

The concept of Legal Protection Debt and Legal Protection by Design might be distant from the perception of liability that characterises current ML-practices. In this sense, it is necessary to add some considerations with respect to the "handoff" approach discussed above. First, such approach has arguably proven to be inefficient and results into a lose-lose scenario. The – technical, ethical and legal protection - debt accumulated along the ML-pipeline tends to be handed off downstream to providers or users of ML systems[39]. This might result might into a downstream legal compliance bottleneck. In order to place on the market, put into service, or use, ML systems, providers and users must satisfy a set of legal requirements. Where the debt that weighs on such stage of the pipeline is very high, providers and uses are faced with the following options: i) they can pay off the debt, incurring in a potentially severe cost; ii) in some cases, they might abandon their project, due to either the too high cost or the impossibility to solve the debt; iii) they might take the risk, handing-off the debt accumulated to the next actor downstream. In the latter case, or whenever, negligently or intentionally, the technical and legal protection debt is not settled by the last actor of the pipeline, the ultimate payer of the debt are the natural persons whose rights and freedoms are affected by the use of the ML-system. Secondly, whatever the perception of ML practitioners, in many cases, the practice of handing-off responsibility downstream the pipeline might not be allowed by the current and the incoming legal framework[40]. This report shows that, when dataset creation, curation and dissemination involve the processing of personal data, the GDPR establishes a framework that imposes practitioners to proactively address the accumulation of legal protection debt along the entire ML pipeline.

The report aims at illustrating how data protection law lays down a set of legal requirements that overcome modularity and encompass the pipeline in its entirety, connecting the actors upstream with

[37] Catherine Petrozzino, 'Who Pays for Ethical Debt in AI?' (2021) 1 AI and Ethics 205. https://doi.org/10.1007/s43681-020-00030-3

[38] Mireille Hildebrandt, 'Saved by Design? The Case of Legal Protection by Design' (2017) 11 NanoEthics 307. Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* (Edward Elgar Publishing 2015).

[39] Widder and Nafus (n 26).

[40] In this sense, the proposed AI Act establishes obligations that apply mainly to actors downstream the pipeline, i.e., use, making available on the market and putting into service of a high-risk AI system. However, the AI Act also establishes duties of care of data governance that might be have upstream effects. Similar considerations apply with respect to the proposed recast of the directive on defective product liability, with respect to the provisions on the liability of the manufactures of a component of a defective product. See, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, art. 10(2) to (4); Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM/2022/496 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496, art. 7, § 1; art. 10, § 1, f.

those downstream. The GDPR establishes rules relating to the protection of natural persons with regard to the processing of personal data[41]. The protection offered by the GDPR covers fundamental rights and freedoms of natural persons "*and in particular* their right to the protection of personal data"[42]. This means that the GDPR is not just about the protection of privacy. The GDPR governs the processing of personal data in order to safeguard all rights and freedoms that can be affected, directly or indirectly, by the processing. The addressees of protection are natural persons, and "not just" data subjects.

In order to ensure the effectiveness of protection, the framework established by the GDPR distributes responsibility among all the controllers that processes personal data, whatever their position in the pipeline. The GDPR makes controllers responsible for the effects of the processing that they carry out. As we will argue in section 2, this means that, whenever controllers operate within a broader pipeline, they are asked to take into consideration also the effects on rights and freedoms of natural persons that the datasets they create or curate might produce in the downstream stages of the pipeline. In this sense, we will show how the GDPR provides the tools to mitigate the problem of many hands in ML-pipelines. The duties and obligations set by the GDPR require controllers to implement by design safeguards that conjugate the need to address downstream harms with the necessity to comply with the standards that govern scientific research. In this perspective, we claim that the obligations established by data protection law either instantiate or harden most of the requirements set by the Open science and Open data framework and also the best practices emerged within the ML-community.

In the next section we set the stage for the analysis of how data protection law applies to dataset creation, curation and release. To this end, we briefly introduce a set of fundamental concepts of data protection law. We then illustrate how such concepts apply in the context of a modular, multi-actor, ML-pipeline.

1.2. The Dataset pipeline and data protection *in itinere*

The GDPR establishes a normative framework that aims at ensuring the protection of personal data and the rights and freedoms of natural persons that are affected by data processing. Such goal is pursued by establishing norms that govern the processing of personal data for their entire life cycle. The GDPR defines "processing of personal data" as any operation performed on personal data[43]. As we will illustrate more in details in section 2, the system of protection set by the GDPR hinges upon the attribution of responsibility to the subject that determines the means and purpose of the processing, i.e., the data controller[44]. The obligations that controllers are required to comply with varies depending on, inter alia, the nature, purposes and context of the processing. In order to analyse how such obligations apply to those processing activities that are involved in the creation, curation and release of datasets, it is necessary to draw some terminological distinctions. Based on the terminology employed by the GDPR[45], we distinguish between two cases:

i) *collection* of data, or self-collection, i.e., the cases in which controllers collect data *directly from the data subject*;

ii) *obtaining* of data, i.e., the cases in which controllers obtain data *from sources other than the data subject*. In this case, the controller processes an already existing dataset that can be or not publicly available, as in the case of obtaining of off-the-shelf datasets.

In fig. 3 we offer a preliminary overview of how the above working definitions apply to the activities of creation, curation and release of datasets in a hypothetical simplified pipeline.

Figure 2: A simplified ML-pipeline and data protection law

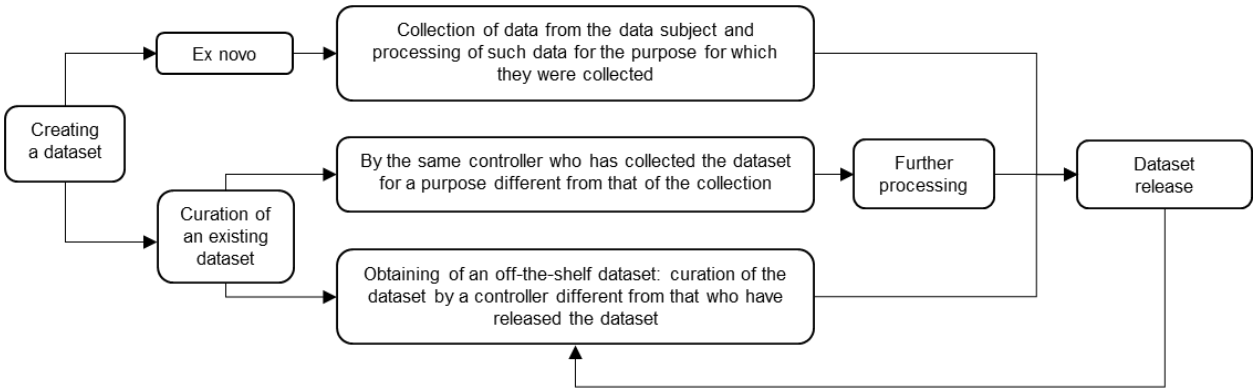
[41] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), http://data.europa.eu/eli/reg/2016/679/2016-05-04 (hereafter GDPR), art. 1, § 1

[42] GDPR, art. 1, § 2, emphasis added

[43] N.B.: *any* operation, including the mere storage and even the erasure or destruction.

[44] Art. 4, § 1, n. 7, GDPR. Controllers can avail themselves of data processors, i.e., the subjects who process data on behalf of the controller (Art. 4, § 1, n. 8)

[45] Art. 13 and art. 14 GDPR. N.B.: this distinction is not used consistently in the text of the GDPR.



In the box below we illustrate the different hypotheses represented in the figure.

1. Dataset creation by *self-collection*
The controller creates the dataset, as it were, *ex novo*. The controller directly collects the data from the data subjects and process the data for the purpose for which they were collected, producing a dataset.

2.Creation by curation
The controller curates an existing datasets, creating a new dataset. We distinguish two cases:

2.1. "In-house" curation of a self-collected dataset
The controller has collected the data directly from the data subjects and has produced a dataset. Later, the controller further processes such data with the purpose of creating a new dataset. A difference must be struck between:

2.1.1. in-house curation for the same purpose or for a purpose compatible with the purpose for which the data were initially collected

2.1.2. in-house curation for a purpose different/incompatible with the purpose for which the data were initially collected.

2.2. Curation of an *obtained* dataset
The new controller obtains a dataset directly (e.g., data transfer) or indirectly (i.e., from a repository) from the old controller and curates such dataset. The curation of an obtained dataset can be further distinguished between the following cases:

2.2.1. the old controller/controller *a quo* from which the dataset has been obtained have self-collected the data directly from the data subjects;

2.2.2. the old controller/controller *a quo* from which the dataset has been obtained has herself obtained the dataset from another controller.

3. Release of a dataset
The controller discloses a dataset. This may include the transmission of the dataset to an identified recipient as well as the dissemination or otherwise making available of the dataset by publication, deposit in a repository, etc..

Each of the activities performed by controllers in the cases illustrated above is governed by a set of general provisions set by the GDPR, e.g., those established in Chapter 1, 2 and 4 of the GDPR. At the same time, the performance of the activities distinctive of each case can trigger the application of specific legal provisions, e.g., the requirements that apply only in case of transfer of personal data. In section 2 we will examine how a set of requirements "follow" datasets in the various phases of processing, obliging controllers to the assess the processing activities that they perform in the present in the light of potential downstream processing. However, before delving into the analysis of the specific requirements set by the GDPR, it is necessary to briefly examine the conditions of applicability of the legal regime established by data protection law.

1.2.1. What kind of data are collected, obtained, released?

In collecting, obtaining or releasing a dataset, the first and crucial question that controllers must answer is what kind of data are processed. Data protection law, and the GDPR, applies only to the processing of personal data, that is, any information concerning an *identified* or *identifiable* natural

person[46]. The assessment of the identifiability of a natural person must be grounded on the consideration of "all the means reasonably likely to be used" - from the perspective of both the controller and a possible future controller – "to identify the natural person directly or indirectly"[47]. By making reference to a result – the identification – or the possibility of such result – the identifiability – the definition of personal data is context-dependent. The qualification of data as personal or non-personal can change as a result of the processing to which the data are subject and external circumstances. This imposes a precautionary approach according to which, in doubt, data should be treated as if they were personal[48].

When collecting, obtaining or releasing a dataset containing personal data, a further step is the identification of the category to which such personal data belong. The most relevant categories of data are the following:

- special categories of data whose processing is prohibited by art. 9 and 10: before collecting or obtaining data belonging to these categories, controllers must make sure that they can rely on a lawful ground for processing, as provided by art. 9, § 2, and art. 10 GDPR.
- pseudonymised data. The GDPR defines pseudonymisation as that processing of personal data that results in the personal data being no longer attributable to a specific data subject "without the use of additional information"[49]. For the data to be considered pseudonymised, such additional information must be "kept separately" and must be "subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person"[50]. As we discuss more in detail in the Annex, pseudonymisation represents a measure that assumes particular relevance under the principle of data minimisation and especially under the special regime that applies to processing performed for research purposes. In any case, it is important to stress that pseudonymised data are still personal data[51]. This implies that both the controller who has performed the pseudonymisation processing[52] and the eventual further controller who obtains off-the-shelf pseudonymised data are still obliged to comply with the requirements set by the GDPR. At the same time, under certain conditions, the processing of pseudonymous data can allow the lessening of controllers' obligations to comply with data subjects' rights (e.g., art. 11 GDPR).

Figure 3: Anonymisation and Pseudonymisation[53]

[46] Pursuant to art. 4, § 1, 1, GDPR, 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

[47] In order to assess what are the means of identification "reasonably likely to be used, Recital 26 prescribes an "objective test". According to such test, based on the consideration of the available technology and its developments, the controller must assess the cost and amount of time that would be required for the identification. Finck and Pallas distinguish between and discuss the "no-risk approach" endorsed by the Article 29 Working Party and the "risk-oriented approach" enshrined by Recital 26: see 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; Article 29 Working Party, Opinion 05/2014 on Anonymisation Techniques, Adopted on 10 April 2014, 0829/14/EN, WP216, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[48] Cf. 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, http://data.europa.eu/eli/reg/2018/1807/oj, art. 2, § 2, second paragraph, according to which "Where personal and non-personal data in a data set are inextricably linked, this Regulation shall not prejudice the application of Regulation (EU) 2016/679. Cf also Communication from the Commission to the European Parliament and the Council, Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union, COM/2019/250 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN .

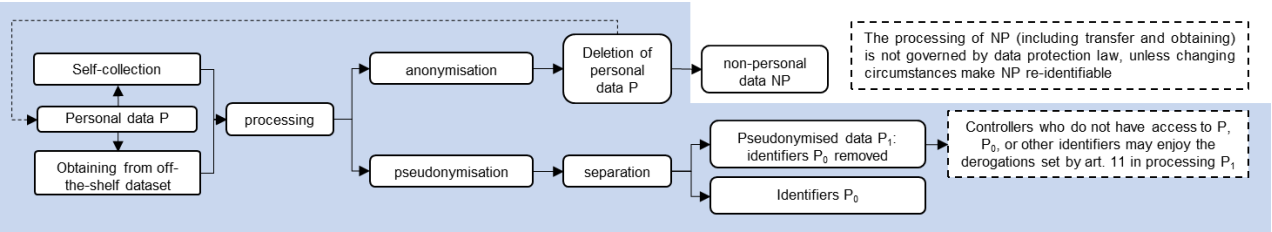
[49] Art. 4, § 1, n. 5, GDPR

[50] Ivi

[51] Recital 26, GDPR

[52] Pursuant to Recital 29 GDPR: "… measures of pseudonymisation should, whilst allowing general analysis, be possible within the same controller when that controller has taken technical and organisational measures necessary to ensure, for the processing concerned, that this Regulation is implemented, and that additional information for attributing the personal data to a specific data subject is kept separately. The controller processing the personal data should indicate the authorised persons within the same controller".

[53] The area in blue encompasses the processing of *personal* data. Such area, therefore, corresponds to the scope of application of the GDPR.



1.2.2. The GDPR and the special research regime.

In the following sections, the report will examine the application of the GDPR to the activity of creation, curation and release of datasets in the context of scientific research. Section 2 will address the core non-derogable obligations that characterise the system of protection established by the GDPR. In the Annex, we will analyse the special regime provided by the GDPR for processing of personal data is carried out for scientific research purposes. Before analysing the general obligations that inform controllers' responsibility under the GDPR, it is worth to briefly delineate the scope of the special research regime:

1. Principle of segregation: the derogations provided by the research regime find application *exclusively* to those data processing activities that are carried out for research purposes. Parallel or subsequent processing activities which pursue, also partially, a purpose different than scientific research do not benefit from the derogations.

2. All derogations are "under condition": the fact that a certain processing activity is performed for scientific research purpose is a necessary but not sufficient condition to trigger the applicability of the derogatory provisions set by the GDPR. Most of the derogations find application only when i) data are processed (exclusively) for research purposed *and* ii) the other conditions set by art. 89 and the single provision in question are met. There is no automatism of the form "research purpose = applicability of the derogation". Unless all the conditions necessary for the application of the derogation are satisfied, the GDPR applies in full.

3. The scope of the regime of derogations is limited: most of the provisions contained in the GDPR are not subject to any derogation or exemption in view of the scientific research purpose of the processing. All in all, the research regime provided by the GDPR covers the application of a limited number of provisions (or part of provisions). A processing that is unlawful in that it does not comply with the general provisions set by the GDPR cannot enjoy the effects of the derogations provided by the research regime. This is a non-exhaustive list of obligations which are not subject to derogation under the research regime:

- Obligations descending from the principles relating to processing of personal data (art. 5):
  - lawfulness, fairness and transparency (5, 1, a);
  - purpose specification (5, 1, b)
  - data minimisation (5, 1, c);
  - data accuracy (5, 1, d);
  - data integrity and confidentiality (5, 1, f);
  - accountability (5,2).
- Obligation to ground data processing on a valid legal basis (art. 6)
- Requirements of the consent (art. 7)
- General provision related to the rights of the data subject (art. 12)
- General responsibility of the controller (art. 24) and joints controllers (art. 26)
- Obligation to implement data protection by design and by default (art. 25)
- Obligations concerning data processors (art. 28)
- Obligation to keep records of processing activities (art. 30)
- Obligation to adopt technical and organisational measures to ensure the security of the processing (art. 32)
- Obligation to foresee risks and, in some circumstances, to make a data protection impact assessment (art. 35) and proceed to prior consultation with the Data Protection Authority (art. 36)
- Obligation to have in place technical and organisational measures to communicate a data breach to the data subject and/or to notify the data protection authority.

- Obligation to designate a data protection officers (art. 37)
- Obligations concerning the transfers of personal data to third countries or international organisations (art. 44-49)

Such a preliminary overview of the scope of the special research regime helps highlighting a circumstance that assumes particular relevance to understand the rationale of the regime of responsibility established by the GDPR. The derogations allowed under the special research regime concern almost exclusively the GDPR provisions on the rights of data subjects, while no derogation is possible for the obligations listed above. The derogations provided under the special research regime, as well as those provided by art. 11 in case of strong pseudonymisation, allows controllers to modulate their obligations towards data subjects where the processing of personal data is not likely to affect significantly the natural persons that are identified or identifiable through such data. As it were, the decrease of the level of potential harm makes possible the lessening of the safeguards required to ensure the protection of data subjects. Even in such cases, however, no derogation is allowed with respect to the requirements different than those concerning the rights of data subject. This circumstance makes manifest that the system established by the GDPR aims at providing a form of protection that goes beyond the natural persons whose personal data are processed at that time by controllers. As we discuss in the following section, the GDPR institutes a regime of responsibility under which controllers are called to consider and address the broader implications of their processing activities.

2. Addressing Legal Protection Debt through Legal Protection by Design

In Section 1 we have briefly illustrated how some of the structural features of ML-pipelines favour a siloed perception of responsibility. We have highlighted how such approach to responsibility can spread especially among the actors upstream the pipeline. In this section we contrast such siloed approach by illustrating how the GDPR establishes a framework of legal protection that does not admit the modularisation of responsibility. Data protection law requires controllers to constantly assess how the decisions they take upstream affect the likeliness and severity of risks downstream. We show how the correct interpretation of the provisions established by the GDPR requires controllers to adopt compliance practices that can address the accumulation of legal protection debt.
In section 2.1. we illustrate the core structure of the regime of liability to which controllers are subject under the GDPR. We show how such regime of liability hinges upon controllers' duty to perform a context-dependent judgment. Such judgment must informs controllers' decisions as to the measures to be adopted to ensure compliance with all the obligations established by the GDPR. We then examine how such judgment must be based on the consideration of the downstream harms posed by the processing.
In section 2.2. we show how the compliance with the documentation obligations set by the GDPR can mitigate the accumulation of a documentation debt and ensure controllers' compliance with the obligations established by other normative framework, such as Open Data and Open Science.
In section 2.3. we analyse the requirements that govern the release and downstream (re)use of datasets.

2.1. Liability and risk under the GDPR

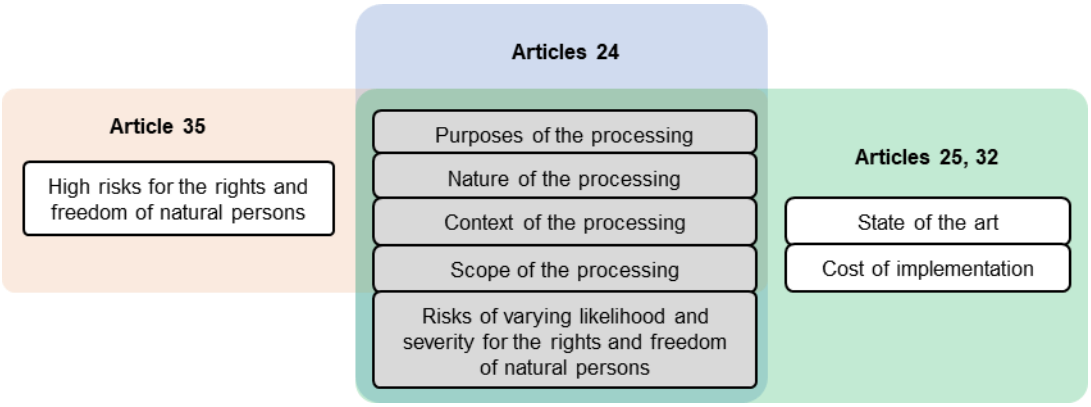
The GDPR subjects controllers to liability for the harm caused by the processing of personal data. In particular, controllers must compensate "any person who has suffered material or non-material damage as a result of an infringement of [the GDPR]"[54]. This regime of civil liability applies to all data controllers, whatever their position – upstream or downstream – in the ML-pipeline.
The complexity of the ML-pipeline – e.g., how "many hands" are involved - and the position occupied by the controller - more or less proximate to the harm, may affect the ascription of liability for damages to controllers. The establishment of controllers' liability for damages is less complex in cases of, as it were, "controller-wise integrated pipelines". By this, we mean the cases in which the same controller creates/curates a dataset, train a ML model on the dataset, implements such model into a system and uses the system in a way that causes harm to natural persons[55]. In this scenario, the misconduct of the controller causes the suffering of damages. The addressee of the duties of care imposed by the GDPR coincides with the subject who has caused the damage. However, as discussed in section 1, in the current ML practice it is not unlikely that controllers who create, curate and release datasets do not engage in the further stages of the ML pipeline. The attribution of liability for damages becomes more complex where more "hands" are in between the controller that creates or curates a dataset and the downstream processing of personal data that eventually harms natural persons. In such cases, the harm might be caused by other controllers that have curated the parent dataset, further controllers that have then used such dataset to train a ML, further downstream actors that have implemented such model into a system, and/or further distant actors that have used such system to make decisions affecting natural persons. The identification of a legally relevant link between the processing performed by the different controllers and harm suffered by natural persons is more straightforward for what concerns the controllers downstream, i.e., those controllers who perform a processing that is more proximate to the harm. Where the pipeline is more complex, the link between the processing carried out by the controllers upstream and the harm suffered downstream might be lessened. Depending on the different norms on negligence and causation across Member States, it is possible that the actions of downstream controllers might interrupt the link between the misconduct of controllers who create, curate and release a dataset upstream and the harm suffered downstream[56].

[54] Art. 82 GDPR
[55] Cf section 1.1., figure 1
[56] Cf, *mutatis mutandis*, the provisions on the effects of substantial modification of a product or AI-system: Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial

However, this does not translate into a lack of liability on the part of upstream controllers. The GDPR establishes a framework under which controllers must take responsibility for "what happens downstream" even where the rules governing civil liability might not apply. All controllers, whatever their position in the pipeline, are required to comply to the duties of care set by the GDPR and are responsible for the consequences of their violation. The infringement of such duties of care can make controllers subject to corrective measures[57] and "effective, proportionate and dissuasive" administrative fines by Data Protection Authorities[58]. Such corrective measures can consist also in a temporary or definitive limitation or a ban on processing[59].

The responsibility framework established by data protection law hinges upon controllers' duty to implement measures to *ensure* and *be able to demonstrate* the compliance of processing activities with the GDPR[60]. Controllers are required to recursively perform an *assessment of appropriateness and effectiveness* of the measures adopted - and to be adopted - to ensure compliance with the GDPR. The performance of such assessment requires controllers to take into consideration multiple factors. While the scope of such factors can be enlarged depending on the measures under consideration[61], art. 24 GDPR identifies the core factors that must inform controllers' assessment: the nature, scope, context and purposes of the processing and, the risk to the rights and freedoms of natural persons posed by the processing.

Figure 4: Factors to be assessed by controllers under the GDPR



The consideration of the abovementioned factors guides controllers in the identification of the content of the obligations that they are required to comply with under the GDPR. As it were, the assessment of such factors aims at answering the question "What am I expected to do in order to comply with the GDPR?"[62].

It should be noted that the factors are intrinsically intertwined and, ultimately, they all affect the assessment of the risk posed by the processing. In the architecture of the GDPR, the concept of risk is an instrument to assess the content of obligations in light of the principle of precaution. In this

---

Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, art. 3, § 1, n. 23, art. 28; Proposal for a Directive of the European Parliament and of the Council on liability for defective products, COM/2022/495, final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A495%3AFIN, Recital 29; art. 4, § 1, n. 5, b), art. 7, § 4, art. 10.

[57] Art. 58, § 2, GDPR

[58] Art. 83 GDPR. Where provided so by Member State law, the violation of the GDPR can also give rise to criminal liability, see art. 84 GDPR

[59] Art. 58, § 2, f, GDPR

[60] Art. 5, § 2, art. 24, R 74, GDPR

[61] See, figure 4. For instance, the text of the provisions on Data Protection by Design (art. 25) and the security of processing (art. 32) adds a reference to "the state of the art" and "the cost of implementation".

[62] A certain kind of processing (nature) involving a limited number of subjects (scope) carried out for a certain *purpose* in certain circumstances (context) might give rise to limited *risks*. In such cases, controllers "may not have to do as much to comply with its legal obligations as a data controller whose processing is high-risk", Article 29 Working Party, Statement on the role of a risk-based approach in data protection legal frameworks, Adopted on 30 May 2014, 14/EN, WP 218, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf?wb48617274=72C54532, p. 2

sense, the GDPR adopts a risk-based approach[63]. Under such approach, the concept of risk play a role that goes beyond that of the specific tools of the data protection impact assessment[64]. The risk that assumes relevance under the GDPR is the risk of a harm to the rights and freedoms of natural persons[65].

This means that the potential harms that the processing may cause to natural persons must inform all controllers' decisions on what measures are necessary to ensure their compliance with the GDPR. Accordingly, the decision as to whether controllers' conduct respects or violates the requirements set by GDPR also depend on the consideration of the potential harms caused by the processing. Where controllers fail to consider the foreseeable risks posed by their processing, they lack the knowledge basis to determine which measures would constitute an adequate and effective implementation of the duties imposed by the GDPR. This can lead controllers to adopt inadequate and ineffective measures – or none - thereby failing to comply with the requirements established by data protection law.

The scope of risks that controllers are expected to foresee depends on multiple circumstances. It is important to highlight that the scope of potential harms that controllers are required to address encompasses also the natural persons that are not in a relation of immediate proximity with the processing carried out by the controller, i.e., also the natural persons that are not, or not yet, data subjects.

A first set of factors that determine the area of the foreseeable risk is provided by the abovementioned elements indicated by art 24, i.e., the nature, scope, context, purposes of the processing. The foreseeability of risks also depends on the personal capacities of controllers, such as their knowledge and skills. In this sense, it has to be noted that controllers who create, curate and release datasets in the context of scientific research are in a qualified position. Such qualified position affects the scope of the risks that controllers-researchers can be expected to foresee. Controllers are also expected to be aware the structural features ML-pipelines discussed in section 1. Dataset practitioners cannot miss to consider that datasets are artifacts infrastructural character[66]. Datasets are created, curated and released to make possible the performance of a set of activities, from further curation to the training and testing of ML models, to the implementation of such models into decision-making systems. At the same time, controllers cannot avoid to consider that datasets can be re-used for scientific research as well as commercial purpose, and also for the performance of public tasks. The scope of the risks that controllers can be expected to foresee depends also on the purposes for which datasets have been created, curated and released. The scope of foreseeable risks covers, as a minimum, the risks connected to the use-cases inherent to the purposes for which the dataset has been created, curated, released. Especially when a dataset is meant to be made publicly available and re-usable, a great share of "what happens downstream" is expectable, to the extent that downstream uses reflect precisely what the creation or curation of a dataset aimed at. In this sense, it is worth stressing that the scope of the foreseeable downstream risks posed by the processing can depend on, but is not limited by, the purpose of the processing. The foreseeability of "what might happen downstream" also depends on the approach adopted by controllers with respect to the dissemination and re-use of datasets. The less safeguards are taken in this sense, the broader the spectrum of possible harms that may occur. Within such spectrum, a set of risks firs within the scope of the foreseeable risk. To some extent, the future uses of a dataset are inherently open-ended - and, as all technological artifacts, datasets are multi-stable[67]. However, datasets are not created, curated and released into a vacuum. The standards of foreseeability that apply to controllers are determined by the latter's expertise in the specific context of use for which datasets have been designed as well as the awareness of broader ML practices. The belonging to the ML community, together with the

[63] Ibid.; Gellert, Raphaël, The Risk-Based Approach to Data Protection (Oxford University Press, 2020); Katerina Demetzou, 'GDPR and the Concept of Risk': in Eleni Kosta and others (eds), *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data: 13th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Vienna, Austria, August 20-24, 2018, Revised Selected Papers* (Springer International Publishing 2019) <https://doi.org/10.1007/978-3-030-16744-8_10> accessed 29 April 2023. Claudia Quelle, 'The 'Risk Revolution' in EU Data Protection Law: We Can't Have Our Cake and Eat It, Too': in Ronald Leenes, Rosamunde van Brakel, Serge Gutwirth, and Paul De Hert (eds.), *Data Protection and Privacy: The Age of Intelligent Machines*, (Hart Publishing 2017); Niels van Dijk, Raphaël Gellert and Kjetil Rommetveit, 'A Risk to a Right? Beyond Data Protection Risk Assessments' (2016) 32 Computer Law & Security Review 286. http://dx.doi.org/10.1016/j.clsr.2015.12.017

[64] Art. 35 GDPR; European Data Protection Board, Guidelines on Data Protection Impact Assessment (DPIA), Adopted on 4 April 2017, As last Revised and Adopted on 4 October 2017, 17/EN, WP248rev.01, https://ec.europa.eu/newsroom/just/document.cfm?doc_id=47711

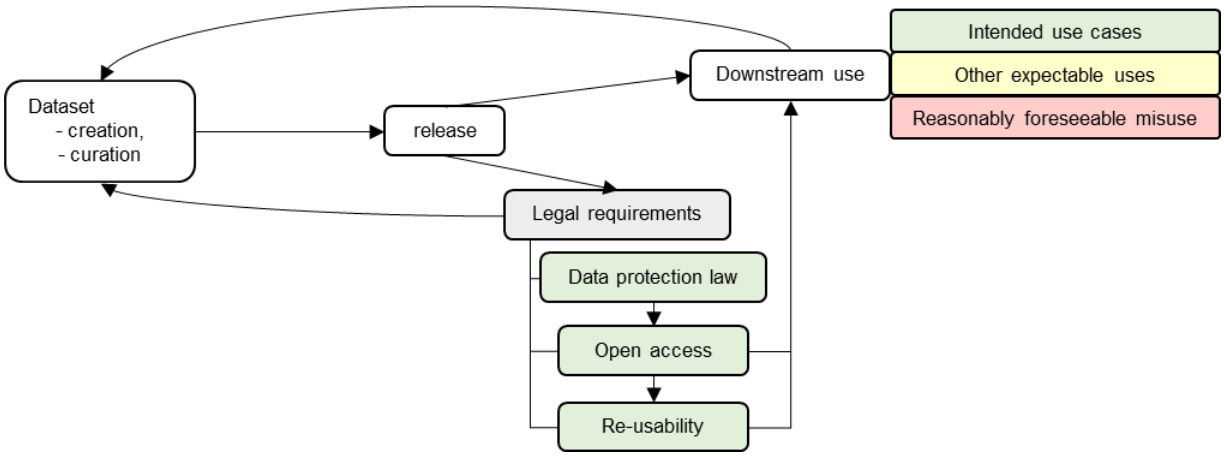
[65] Art. 1, § 2; Recital 75, GDPR.

[66] Hutchinson and others (n 13).

[67] Don Ihde, Technology and the Lifeworld. From Garden to Earth (1990 Indiana University Press)

experience accumulated and best practices emerged within such community, provide researchers with the elements necessary to perform an assessment of what risks are reasonably foreseeable. The assessment of the purpose and context of the processing in light of the experience accumulated within the scientific community can also attribute legal relevance to risks which are in a weaker relation of proximity to the processing performed by a data controller. More and more risks are likely to become foreseeable with the growing findings of scientific literature, but also of the legal decisions adopted by DPAs and the judiciary.
The assessment of the potential harms deriving from the processing informs the decision as to what are the appropriate and effective measures that controllers are required to build-in their processing activities to ensure compliance with the GDPRR. The obligation to adopt adequate measures is a "best effort" obligations. The assessment of what measures fulfil such standard is governed by the same criteria that characterise the assessment of risk.
In essence, the duty to anticipate and address potential downstream harms requires controllers to adopt a forward-looking approach. In order to ensure compliance with the GDPR, controllers must engage in a dynamic, recursive practice that addresses the requirements of present processing in the light of the future potential developments. At the same time, the planning effort required by the GDPR is strictly connected with the compliance with obligations set by other normative frameworks. In this sense, compliance with the GDPR and compliance with obligations such as those imposed by the Open science and Open data framework go hand in hand. As we will see in the following sections, compliance with the GDPR is a pre-requisite for complying with Open science and Open data framework. Simultaneously, the perspective of open access and re-usability of datasets affects the content of the obligations set by the GDPR.
As a result, the consideration of "what happens downstream" - i.e., the potential uses of datasets, potential harms that the latter may cause, further requirements imposed by other normative frameworks – backpropagates, determining the requirements that apply upstream.

Figure 5: "backpropagation" of the requirement that apply to the downstream stages of the pipeline



2.2. Documentation

In section 1 we have seen that the recent literature has emphasised how current practices often lead to the accumulation of a documentation debt along the ML-pipeline[68]. We have also emphasised how documentation debt can turn into a legal protection debt. In this section, we show how the implementation of adequate documentation practices can contribute to ensure that legal protection concerns are addressed both upstream and downstream the ML-pipeline. Documentation practices favour accountability by requiring controllers to both *take into account* potential downstream harms and *give an account* of the decisions taken to address such harms[69].

[68] Section 1.1., n 36
[69] Karen L Boyd, 'Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data' (2021) 5 Proceedings of the ACM on Human-Computer Interaction 438:1.https://doi.org/10.1145/3479582

The production and keeping of documentation on the choices made in designing a dataset assumes particular relevance in the context of scientific research. For the purposes of scientific research, the production of documentation concerning the dataset is arguably as important as the dataset itself. Documentation constitutes in this sense a requirement functional to the dissemination of research and the re-use of its output. In this perspective, the Commission Recommendation on access to and preservation of scientific information welds the concept of open access with the need to adopt data management planning practices since the early stages of generation and collection of data in the context of the research process[70]. The Horizon Regulation consolidates the bond between Open science and responsible management of research data by requiring beneficiaries of research grants to establish a Data Management Plan[71]. Moreover, the keeping of documentation tools such as Data Management Plans, or Data Protection Concepts, constitutes a requirement that the law of some Member State imposes to researchers prior to the start of, and in the course of, research projects[72].

In line with the principle of accountability, the GDPR contains several provisions that impose controllers to comply with documentation requirements. We claim that compliance with such documentation requirements can perform a double function, i.e., prevent the rise of legal protection debt and contribute to the meeting of the standards that govern scientific research.

As discussed above, under the GDPR the controller is *responsible for*, and *shall be able to demonstrate,* compliance with data protection law[73]. Therefore, the production and keeping of documentation assumes relevance for both compliance and the capacity to demonstrate compliance. In some cases, keeping documentation is in itself the content of an obligation, e.g., the obligation to keep record of processing pursuant to art. 30 GDPR. In other cases, the keeping of documentation is a necessary pre-condition for complying or demonstrating compliance with other obligations. For instance, compliance with the obligations to provide information to data subjects[74] requires, first, that the controller has duly collected the information that she is required to provide; secondly, that the controller keeps track of, i.e., documents, having provided such information. Even when the keeping of documentation is not directly prescribed by the GDPR, or it is not immediately instrumental to demonstrate compliance with specific obligations, documentation practices can promote reflexive practices[75]. Such practices can facilitate the assessment of adequacy and effectiveness that, as discussed in the previous section, is at the core of compliance with the GDPR.

Without claiming to be exhaustive, in the first column from the left of table below we have attempted to map the main documentation obligations set by the GDPR. The second and third columns from the left juxtapose the documentation requirements established by the GDPR with the requirements that oft-cited data science literature has identified as necessary to avoid documentation debt[76]. In the fourth column, we have included reference to documentation requirements provided by the FAIR

[70] Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information, C/2018/2375, http://data.europa.eu/eli/reco/2018/790/oj, Recital 5, §§ 3, 4

[71] Regulation (EU) 2021/695 of the European Parliament and of the Council of 28 April 2021 establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination, and repealing Regulations (EU) No 1290/2013 and (EU) No 1291/2013, http://data.europa.eu/eli/reg/2021/695/oj (hereafter, Horizon Regulation), art. 14, § 3; art. 39, § 4. EU Grants: Data Management Template (HE):V1.0 – 05.05.2021, https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/reference-documents;programCode=HORIZON

[72] European Data Protection Supervisor, Study on the appropriate safeguards under Article 89(1) GDPR for the processing of personal data for scientific research, Final Report EDPS/2019/02-08, August 2021, https://edpb.europa.eu/system/files/2022-01/legalstudy_on_the_appropriate_safeguards_89.1.pdf , p. 54

[73] Art. 5, § 2, and art. 24 GDPR

[74] Art. 13, 14, 15, GDPR

[75] Cf., *mutatis mutandis*, Miceli and others (n 37).

[76] The table shows questions formulated by Gebru et al. (n 36) and by Pushkarna et al. (n 36). To parity of authoritativeness with other sources in the field, we have included in the table only these two papers out of reasons of practicality, i.e., possibility to easily number, reference and include the questions in the table. In making this choice, we have taken into consideration the fact that there was a significant overlapping between the sources selected and the other sources referred to in the Report. In particular, our decision was based on the consideration that the two articles referenced in the table are overall representative of the concerns and suggestions advanced in the relevant literature. Cf, sources quoted at nn 13, 16, 36.

principles[77]. In the last column on the right we have referenced the requirements set by the Data Management Plan template provided under the Horizon Europe program[78].
The table aims at showing the convergence between the documentation requirements established by normative frameworks that are relevant in processing personal data in the context of scientific research: data protection law, the standards emerging within the scientific community, and the consolidated scientific research requirements set within Open data and Open science framework.
The overlapping between the documentation requirements established by such different frameworks shows firstly that a serious approach to the compliance with the GDPR can provide the safeguards necessary to avoid the accumulation of a documentation debt. In this way, compliance with the documentation obligations set by the GDPR can prevent the accumulation of other forms of technical debt and, eventually, of legal protection debt. At the same time, the convergence between the requirements set by the GDPR and those established by the FAIR principle and the Horizon DMP template shows how the performance of the documentation obligations established by the GDPR can also facilitate compliance with requirements specific to data processing conducted in the context of scientific research.
At this point, it is necessary to stress that convergence between the documentation requirement set by the different framework considered does not result in a complete overlapping. This makes it interesting to look at all the cases in which the requirements identified in sources other than the GDPR are broader or more specific than those provided by the latter. Beside the cases in they are binding by virtue of distinct legal sources[79], the requirements indicated in the four columns on the right can contribute to increase the granularity of the documentation obligations set by the GDPR. We emphasise once again that the assessment of the measures necessary to ensure and demonstrate compliance with the obligations set by the GDPR is centred on the performance of a context-dependent judgment of adequacy and effectiveness[80]. In assessing the adequacy of the documentation produced to ensure compliance with the GDPR, controllers cannot ignore the best practices emerging from the scientific community they belong to. Such best practices assume legal relevance through the judgment prescribed by art. 24, 25, 32, 35 of the GDPR. The sources included in the table and those referenced *supra* in the report[81] describe documentation requirements that can become crucial to ensure full compliance with the GDPR. The documentation requirements indicated by such sources are particularly valuable especially for the assessment and mitigation of risks downstream the ML-pipeline[82]. Moreover, the keeping of adequate documentation can facilitate not only controllers' self-assessment of compliance, but also the assessment of compliance by third parties. These can includes auditors, Data Protection Authorities or other public bodies. Facilitating compliance assessment by third parties is particularly important especially in the perspective of dissemination and re-use of datasets. The provision of adequate documentation can help the recipients of datasets to avoid form of dataset re-use that might result in unintended consequences. Once again, it is necessary to stress that controllers' present decisions must be informed by the consideration of the downstream potential harms. In this perspective, it is crucial that controllers document the intended and potential uses of a dataset. Next to being necessary for risks assessment and mitigation, providing documentation that accounts for future use of a dataset provides recipients with the information necessary to ensure their compliance with data protection law. Such documentation should attest whether the dataset contains personal data, and whether they are pseudonymised, as well as information about the collection, or obtaining, of the data, the operation of processing they have been subject, etc.. The perspective of dataset re-use, moreover, reinforces the necessity that the dataset designers provide detailed documentation on the limitations of the dataset.

[77] Mark D Wilkinson and others, 'The FAIR Guiding Principles for Scientific Data Management and Stewardship' (2016) 3 Scientific Data 160018. https://doi.org/10.1038/sdata.2016.18
[78] EU Grants: Data Management Template (HE):V1.0 – 05.05.2021, https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/reference-documents;programCode=HORIZON
[79] e.g., where the controller is subject to the obligations set by the Open data directive or is a beneficiary of a grant under the Horizon Regulation. See the requirements indicated in the fourth and fifth column
[80] Art. 5, § 2; art. 24, art. 25; art. 32; art. 35 GDPR. This is particularly the case with respect to the duty to implement measures of data protection by design: art. 25 GDPR includes "the state of the art" among the elements that controllers should take into account in assessing the appropriateness of the measures adopted to ensure compliance
[81] (n 36)
[82] E.g., cf question 40 of Gebru et al. (n 36): "Is there anything about the composition of the dataset or the way it was collected and preprocessed/ cleaned/labeled that might impact future uses?"

Table 1

Documentation required by the GDPR	Measures to address documentation debt suggested in literature		FAIR principles[83]	Horizon Data Management Plan[84]
	Pushkarna et al.[85]	Gebru et al.[86]		
Documentation concerning the subjects actively involved in the processing - Joint controllership[87] - Data processors[88] i) 28, § 1: evidence showing that the controller has used "only processors providing sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject." ii) 28, § 3: legally binding act between controller and processor, attesting the existence of the requirements indicated by letters a-h.	C 1, 2	Q. 2, 3, 4; 24		4
Documentation concerning categories of data[89] and categories of data subject[90] - in the cases provided by art. 9 and 10 GDPR, documentation attesting the applicability of an exception to the prohibition to process special categories of data; - in case of obtaining of data from sources other than the data subject, documentation concerning the source from which the personal data originate and whether the data come from publicly accessible sources[91]. - documentation on the accuracy, integrity and quality of the data[92]	C 5, 6, 7, 8, 9, 10, 17, 18, 19, 20, 21, 22, 23, 25, 26, 27, 28, 29	Q 5-23, 25, 27, 33	I, R	1, 2
Processing, technical aspects[93]	C 17, 18, 19, 20, 21	Q 33, ; 37;	F, A, I, R	
Documentation on the **purpose -** and attesting the **legal basis** - **for each distinct processing activity**[94]. In particular, - if the legal basis is consent: requirements of art. 7[95] - if the legal basis is legitimate interest, identification of the legitimate interest and balancing test[96] - in case of further processing for a purpose other than that for which the personal data were collected or obtained, documentation attesting that the controller has provided the data subject prior to that further processing with information on that other purpose and with any relevant further information. Documentation concerning the compatibility test, in particular: i) consideration concerning the legal basis, ii) conditions of applicability of presumption of compatibility for further processing performed for scientific research purposes[97].	C 11, 12	Q 1; 28, 29, 30		1
Documentation on the **storage period** or on the criteria used to determine that period[98]; documentation attesting the existence of the conditions to apply the derogation provided under the scientific research regime[99]	C 4, 5	Q 34; 50, 53, 54, 55	A, R	4
Automated decision making[100]: documentation concerning the existence of automated decision-making, including profiling and meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.				
Documentation on the **security measures** adopted[101]				5
Documentation necessary to demonstrate compliance with the provisions on **data subjects' rights**[102], in particular - information obligations[103] - documentation showing that the controller has provided data subjects with the information of art. 13 and 14, or that an exception to such obligation applies[104]. Documentation attesting the respect of the requirements on the of timing to provide the information[105].		28		
Documentation on the **scientific research regime** - documentation attesting the exclusive purpose of scientific research[106]. - documentation on the measures and safeguards adopted to comply with art. 89, § 1, GDPR[107] - documentation on the requirements for the application of the exceptions provided by the special regime[108]:		Q 26		all
Documentation related to **data transfer** (transmission, dissemination or otherwise making available of data) **and recipients:** - identity of recipients or categories of recipients of the personal data, in particular recipients in third	C 3, 12	Q 31; 37; 39; 40, 41, 43, 44, 45,	F, A, I, R	2, 6

[83] Wilkinson and others (n 77).
[84] EU Grants: Data Management Template (HE):V1.0 – 05.05.2021, https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/reference-documents;programCode=HORIZON
[85] Pushkarna, Zaldivar and Kjartansson (n 36).
[86] Gebru et al (n 36)
[87] Art. 26 GDPR
[88] Art. 28, §§ 1 and 3, GDPR
[89] Art. 14, § 1, d; art. 15, § 1, b; art. 30, § 1, c, GDPR
[90] Art. 30, § 1, c, GDPR
[91] Art. 14, § 2, f; art. 15, § 1, g, GDPR
[92] Art. 5, § 1, c, d, f
[93] Art. 24, 25, 32, 35 GDPR
[94] Art. 5, § 1, a, b; art. 6; art. 7; art. 13, § 1, c; 14, § 1, c; art. 15, § 1, a; art. 24; art. 25; art. 30, 1, b) ; art. 35, § 7, a, GDPR
[95] Art. 7 GDPR
[96] Art. 13, § 1, d; art. 14, § 1, b; art. 35, § 7, a, GDPR
[97] Art. 5, § 1, b; art. 6, § 4; art. 13, § 3; art. 14, § 4, GDPR
[98] Art. 5, § 1, e; art. 13, § 2, a; art. 14, § 2, a; art. 15, § 1, d; art. 30, § 1, f, GDPR
[99] Art. 5, § 1, e; art. 89, § 1, GDPR
[100] Art. 13, § 2, f; art. 14, § 2, g; art. 15, § 1, h; art. 22, GDPR
[101] Art. 30, § 1, g; art. 32, GDPR
[102] Art. 12-22
[103] Art. 12, 13, 14, 15, GDPR
[104] Art. 13, § 4, GDPR; art. 14, § 5, GDPR. See Annex, section A.2.2.3.1.
[105] Art. 13, § 1; art. 14, § 3, GDPR.
[106] Art. 89, §§ 1 and 4, GDPR. See, infra, Annex, section A.1.
[107] Infra, Annex, section A.1.
[108] Art. 5, § 1, b, e; art. 14, § 5, e; art. 17, § 3, d; GDPR. See Annex, section A.2.2.

countries or international organisations[109]; - existence of legal grounds for international data transfer[110] (e.g., existence or absence of an adequacy decision by the Commission, or in the case of transfers referred to in Article 46 or 47, or the second subparagraph of Article 49(1), reference to the appropriate or suitable safeguards and the means by which to obtain a copy of them or where they have been made available)		46, 47, 48 56		
Documentation on **risk assessment and mitigation**[111] - identification of the risks for the rights and freedom of natural persons, their likelihood and severity. - technical and organizational measures implemented, including data protection by design and by default, eventual data protection policies adopted[112], eventual revision of such measures. - in case of high risks, documentation attesting the performance of a DPIA prior to the processing[113]. Documentation containing, at least a systematic description of the envisaged processing operations, the test of necessity and proportionality of the processing operations in relation to the purposes[114]; Documentation attesting the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with the GDPR taking into account the rights and legitimate interests of data subjects and other persons concerned[115].	C 12, 13, 14	Q 31; 37; 39; 40, 41, 48, 56	F, A, I, R	2, 6

2.3. Addressing downstream risks: release and re-use of datasets

The GDPR aims at ensuring "the free movement of personal data"[116]. To this end, the GDPR establishes rules that guarantee that, wherever the data might go, they are granted the same level of protection. Such rules level the playing field within the EU by setting requirements that apply to all processing of personal data. At the same time, the GDPR provides a set of requirements that apply to extra-EU transfer of personal data. The rules on the movement of data assumes particular relevance in the context of scientific research. Next to the natural vocation of research to internationality, the sharing and re-use of data are promoted under the EU Open science and Open data framework. One of the objectives of the European Union is that of "achieving a European research area in which researchers, scientific knowledge and technology circulate freely, …"[117]. To this end, the EU institutions promote Open science, that is, "an approach to the scientific process based on open cooperative work, tools and diffusing knowledge"[118]. Among the eight ambitions of the of the Open science policy, the EU aims at promoting Open data[119]. A major step in this sense has been made with the recasting of the former Public Sector Information directive[120]. The new Open data directive, in line with previous policy documents, moves from the assumption that open access to research data can help addressing "mounting societal challenges efficiently and in a holistic manner", "enhance quality, reduce the need for unnecessary duplication of research, speed up scientific progress, combat scientific fraud, and … overall favour economic growth and innovation"[121]. To this ends, the Open Data directive establishes a set of minimum rules governing the access and re-use of research data[122]. For the scope of the Directive, research data are "documents in a digital form, other than scientific publications, which are collected or produced in the course of scientific research activities and are used as evidence in the research process, or are commonly accepted in the research community as necessary to validate research findings and results"[123]. While the definition of "re-use" provided by the Directive makes reference only to documents held by public sector bodies and public undertakings, it is possible to assume that, in the context of research, re-use means the use of the

[109] Art. 13, § 1, f; art. 14, 1, e; art. 15, § 1, c; art. 17, § 2; art. 19; art. 30, § 1, d; Chapter V, GDPR
[110] Art. 13, § 1, f; art. 14, § 1, f; art. 15, § 2; art. 30, § 1, e; Chapter V, GDPR
[111] Art. 1, 24, § 1; 25; 32, 35, GDPR
[112] Art. 24, § 2, GDPR
[113] Art. 35, § 1, GDPR
[114] Art. 35, § 7, GDPR
[115] Art. 35, § 7, d, GDPR
[116] Art. 1, § 1, GDPR
[117] Consolidated version of the Treaty on the Functioning of the European Union, http://data.europa.eu/eli/treaty/tfeu_2016/oj, art. 179, § 1.
[118] Regulation (EU) 2021/695 of the European Parliament and of the Council of 28 April 2021 establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination, and repealing Regulations (EU) No 1290/2013 and (EU) No 1291/2013, http://data.europa.eu/eli/reg/2021/695/oj, art. 2, § 1, 5
[119] European Commission, Open Science, https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en
[120] Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, http://data.europa.eu/eli/dir/2003/98/oj; European Data Protection Supervisor, Opinion 5/2018 on the proposal for a recast of the Public Sector Information (PSI) re-use Directive, 10 July 2018, https://edps.europa.eu/sites/edp/files/publication/18-07-11_psi_directive_opinion_en.pdf
[121] 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), http://data.europa.eu/eli/dir/2019/1024/oj (hereafter, Open data directive, or ODD), Recital 27
[122] Ibid., art. 1, § 1, c,
[123] Ibid., art. 2, § 1, 9.

research data for a commercial or non-commercial purposes other than the initial purpose[124]. Pursuant to art. 10, § 2, research data shall be re-usable for commercial or non-commercial purposes when the following conditions cumulatively apply: i) the research data result from publicly funded initiatives; ii) the research data have been made publicly available through an institutional or subject-based repository. The Directive requires Member States to implement such obligations by adopting Open Access Policies addressed to research performing organisations (RPO) and research funding organisations (RFO)[125]. Under such Open Access Policies, RPO and RFO should make publicly funded research data openly available in line with the principle of 'open by default' and the FAIR principles[126].

A correct framing of the practices of dataset creation, curation and release in the context of research requires to make an effort towards the integrity of the legal framework as a whole, taking into consideration the relations between Open data, Open science and data protection law. First, it is first important to stress that compliance with data protection law represents a pre-requisite for the achievement of the goals of Open Data and Open Science framework. Lacking the respect of data protection law, the European Research Area - and a European Data Space[127] - would be built on sand. The open accessibility and re-usability of research data processed in violation of the GDPR would affect the derivate datasets and applications developed downstream. This would lead to an endemic accumulation of legal protection debt, jeopardising rights and freedom of natural persons. Next to the express commitment to the respect of human rights[128], the Open data and Open science framework are governed by the principle "as open as possible, as close as necessary"[129]. In application of such principle, data protection law contributes substantively to the determination of the degree to which data can be either open or close. In particular, the Open Data directive makes clear that its application is without prejudice to data protection law[130]. More specifically, in delegating to

[124] Ibid., art. 2, § 1, 11

[125] Ibid., art. 10, § 1

[126] Wilkinson and others (n 77). At the time of writing, the distribution of responsibility for permitting re-use between researchers, RPO and RFO is affected by a certain degree of uncertainty. This uncertainty is also due to the fact that the Directive has not been implemented yet in several Member States. See, Mirelle van Eechoud, Study on the Open Data Directive, Data Governance and Data Act and their possible impact on research, European Commission Directorate-General for Research and Innovation Directorate A — ERA & Innovation Unit A.4 — Open Science, March 2022, doi: 10.2777/71619, at 23.https://eur-lex.europa.eu/legal-content/EN/NIM/?uri=CELEX:32019L1024. See also Implementation of the Public Sector Information Directive | Shaping Europe's digital future (europa.eu). Similar considerations can be made with respect to the open access obligations set by the Horizon Regulation, which establishes that "Open access to research data shall be the general rule under the terms and conditions laid down in the grant agreement". See, Horizon Regulation (n 118), art. 39, § 3, second paragraph. Also in this case, individual researchers are not the direct addressees of the open access obligations.

[127] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A European strategy for data, COM(2020) 66 final, Brussels, 19.2.2020, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52020DC0066

[128] Recital 71 ODD (n 121),

[129] Art. 10, § 1; Recital 28, ODD (n 121); Horizon Regulation (n 118), art. 39, § 3, second paragraph. Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information, C/2018/2375, http://data.europa.eu/eli/reco/2018/790/oj; Council conclusions on open, data-intensive and networked research as a driver for faster and wider innovation, 9360/15, Brussels, 29 May 2015, https://data.consilium.europa.eu/doc/document/ST-9360-2015-INIT/en/pdf; Council conclusions on the transition towards an open science system, 9526/16, Brussels, 27 May 2016, https://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf

[130] Art 1, § 4; Recital 28 and 53, ODD (n 121). Cf., also, Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information, C/2018/2375, http://data.europa.eu/eli/reco/2018/790/oj, § 3. Art. 1, § 2, h, ODD provides that the directive does not apply to "documents, access to which is excluded or restricted by virtue of the access regimes on grounds of protection of personal data, and parts of documents accessible by virtue of those regimes which contain personal data the re-use of which has been defined by law as being incompatible with the law concerning the protection of individuals with regard to the processing of personal data or as undermining the protection of privacy and the integrity of the individual, in particular in accordance with Union or national law regarding the protection of personal data". The interpretation of such provision has risen several doubts. In this sense, European Data Protection Supervisor, Opinion 5/2018 on the proposal for a recast of the Public Sector Information (PSI) re-use Directive, 10 July 2018, https://edps.europa.eu/sites/edp/files/publication/18-07-11_psi_directive_opinion_en.pdf, §§ 14-16. It is worth noticing that art. 1, § 2, h of the ODD reproposes the exact text of the provision added by Directive 2013/37 to Directive 2003/98 (see, Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information, http://data.europa.eu/eli/dir/2013/37/oj, art. 1, § 1, a, (iii); of Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information; Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, http://data.europa.eu/eli/dir/2003/98/oj. On the interpretation of art. 1, § 1, a, (iii); of Directive 2013/37/EU and its relation with the GDPR, see the considerations made by the Grand Chamber of the ECJ and by the Advocate General Szpunar in: Judgment of the Court (Grand Chamber), Case C-439/19, 22 June 2021, ECLI:EU:C:2021:504, §§ 127-128; Opinion Of Advocate General Szpunar, Case C-439/19, 17 December 2020, ECLI:EU:C:2020:1054, §§ 128-129.

Member States the adoption of Open access policies, the Open Data directive requires that such policies take into account data protection[131]. Similarly, the Horizon Regulation establish that grant agreements must ensure the possibility of exceptions to the general rule of open access to research data based on concerns relating to "data protection rules, privacy, confidentiality …"[132]. The principle "as open as possible, as close as necessary" and the abovementioned provisions makes clear that the obligation to make research data publicly accessible and re-usable apply only with respect to data that already comply with the requirements set by data protection law. The Open Data and Open Science frameworks require controllers to be even more attentive to the requirements set by the GDPR. Not only the need to comply with the obligations deriving from the Open Data and Open Science framework cannot be invoked to justify the making available of research data in contrast to data protection law. By failing to comply with the GDPR, controllers put themselves in the condition of not being able to comply with Open data requirements. In each case in which the lack of compliance with the GDPR depends on inexcusable negligence, it is the conduct of the controller that makes data, as it were, "more close than necessary". In this perspective, what we said above for the obligations set by the GDPR is particularly the case for the requirements set by the Open data and Open science framework: such requirements backpropagates along the dataset pipeline, informing the research designs from its beginning. In order to make data publicly available and re-usable, controllers are required to implement by design all the measures necessary to ensure that the openness of data does not lead to infringements of the rights and freedom of natural persons. This implies making sure that all risks resulting from the processing have been duly addressed. In addition to the requirements examined in the previous sections, the Open data framework requires to pay keen attention to the provisions of the GDPR on the movement of data. Under the GDPR, the movement of data, i.e., the "disclosure by transmission, dissemination or otherwise making available" of personal data, constitutes "processing"[133]. Accordingly, such processing activities must comply with general requirements, including a specified purpose and an appropriate legal basis. Moreover, chapter V of the GDPR provides further conditions for the transfer of personal data to third countries or international organisation. Such conditions apply *in addition* to the other requirements set by the GDPR for intra-EU disclosure of data. This means that the processing consisting in an international data transfer must both satisfy the requirement of a legitimate and specified purpose and a lawful basis of processing and rely upon an appropriate legal ground for the transfer[134].

The adoption of technical and organisational measures to document and monitor data transfers is an indispensable pre-condition for compliance with the obligations triggered by data transfers. This is true firstly with respect to the information obligations towards data subjects[135]. Under the scientific research regime, the GDPR admits a conditioned derogation of the information obligations in cases in which dataset are obtained from sources different than the data subject[136]. It is important to underline that only the obligation to *provide information directly to the data subjects* is subject of potential derogation, not the obligation to be in possession of such information[137]. Controllers' capacity to plan and monitor the ingoing and outgoing flow of data is crucial also to ensure the respect of the requirements concerning the purpose and legal basis of processing. Planning and monitoring data flows is moreover necessary to address the risks connected to the movement of research data and, in particular, those posed by the dissemination and re-use of data. As the dissemination of research

[131] In particular, the ODD requires to take into account the risk of reidentification of data subjects posed by accessibility and re-usability of open data. Art. 10, § 1; Recital 16, ODD

[132] Horizon Regulation (n 118), art. 39

[133] Art. 4, § 1, n. 2, GDPR. Pursuant to art. 4, § 1, n. 9, "recipient" is the subject to which the personal data are disclosed

[134] Chapter V, GDPR

[135] As a general rule, in the case of both self-collection and obtaining of off-the-shelf datasets, controllers are required to provide data subjects with information on "the recipients or categories of recipients" of the data (art. 13, 1, e; 14, 1, e, GDPR). If the recipients are determined, the information obligation must be complied with at the moment of the collection (if the controller collects the data from the data subject) or at the latest when the personal data are first disclosed (art. 14, § 3, c), where the controller obtains the data from sources other than the data subject (art. 14, § 1, e). Where the controller has communicated only the *category* of recipients, the controller must communicate the identity of the actual recipients as soon as the latter are identified. In any case, the controller is required to communicate the identity of the recipients at the request of the data subject pursuant to art. 15 GDPR. In case of international transfer, controllers are required to inform data subjects of the intention "to proceed to international transfer of the data and to provide specific information about the grounds adopted by the controller to legitimise such transfer. In particular, controllers are required to inform data subject i) of the existence or inexistence of an adequacy decision; ii) of the appropriate or suitable safeguards and the means by which to obtain a copy of them or where they have been made available, in the case provided by art. 46, 47, 49, § 1, second subparagraph; iii) etc., see Chapter V, GDPR.

[136] Art. 14, § 5, e, GDPR. See, Annex, section A.2.2.3.1.

[137] Annex, section A.2.2.3.1.

data is a legally qualified goal under the Open data and Open science framework, making datasets containing personal data publicly available is also an activity essentially connotated by risk. It is therefore of the utmost importance that controllers consider nature, scope, context, purpose of the processing-transfer and adopt all the safeguards necessary to ensure legal protection against potential downstream harms. In line with the remarks made in sections 2.1. and 2.2., controllers' assessment of the potential downstream harms must take into account the concerns raised by the ML scientific community. As we have discussed in section 1, the recent scientific literature has addressed the lifecycle of datasets, and especially the life of dataset after their release online. Particular emphasis has been given to the risks posed by "runaway data", i.e., the "phenomenon where data is available through a multitude of sources outside a creator's control"[138]. Once datasets are made publicly available, they can become the starting point of innumerable ML pipelines, or be incorporated into existing pipelines. Without the adoption of by design measures to govern data flows, runaway data can increase exponentially the likeliness and severity of downstream harms.

Compliance with the requirements set by the GDPR is essential to avoid that dataset dissemination gives rise to the accumulation of legal protection debt along ML pipelines. Based on the assessment of adequacy and effectiveness required for all forms of processing, controllers can consider the adoption a range of measures to ensure that data transfer are compliant with the GDPR.

Among such measures, an adequate use of licenses can help restricting unintended and potentially harmful downstream effects[139], implementing the "do no harm' principle[140]. As anticipated in section 2.2., dataset documentation practices are particularly important with respect to data transfers and data re-use. Providing adequate dataset documentation can also reinforce the effects of licenses. Dataset documentation can define in detail the intended uses and users of the released dataset, and exemplify what would constitute an abuse of the dataset. Dataset documentation can also support and justify the limitations on re-use imposed by controllers. For instance, information concerning the quality of data, the modality of collection, the curation processing performed, etc., can provide recipients with further elements to understand why the controller who has released the dataset has considered it unsuited for certain uses.

Further measures are necessary to make sure that the promotion of open access to research data through publicly accessible repositories is compliant with the GDPR. The relationship between the controller and the repository assumes relevance under data protection law. The qualification of the repository as a processor, controller-recipient or joint controller will depend on how powers and responsibilities are in practice articulated between the controller and the repository. This means that both, at the moment of the selection of the repository and for the duration of the relation with the repository, it is important that the controller monitors the respective roles. In most cases, the repository will act as a data processor by carrying out, on behalf of the controller, the processing operations consisting in the "storage" and "making available" of personal data. For this to be the case, controllers must make sure that the repository-processor will provide assistance in ensuring compliance with the GDPR and, among others, delete or return "all the personal data to the controller after the end of the provision of services relating to processing, and deletes existing copies"[141]. More in general, controllers are subject to an obligation to choose only processors that offers organisational and technical measures to ensure compliance with the GDPR. In the context of dissemination and re-use of research, such measures should include adequate access management tools and instrument to ensure the traceability of dataset recipients. Controllers should adopt measures capable of ensuring oversight on, and potentially limit, the access to dataset by recipients. Such measures seems to be an indefectible precondition to ensure compliance with the requirements established by

[138] Kenny Peng, Arunesh Mathur and Arvind Narayanan, 'Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers' (arXiv, 21 November 2021) <http://arxiv.org/abs/2108.02922> accessed 28 April 2023. p. 5; Adam Harvey and Jules LaPlace. Exposing.ai. https://exposing.ai, 2021.

[139] For an analysis of the (in)effectiveness of prevalent licensing practices, see Peng et al., pp. 10-11; See, also, Misha Benjamin and others, 'Towards Standardization of Data Licenses: The Montreal Data License' (arXiv, 20 March 2019) <http://arxiv.org/abs/1903.12262> accessed 28 April 2023. Cf., also, *mutatis mutandis*, European Data Protection Supervisor, Opinion 5/2018 on the proposal for a recast of the Public Sector Information (PSI) re-use Directive, 10 July 2018, https://edps.europa.eu/sites/edp/files/publication/18-07-11_psi_directive_opinion_en.pdf, §§ 33-34.

[140] European Data Protection Supervisor, Opinion 5/2018 on the proposal for a recast of the Public Sector Information (PSI) re-use Directive (n 139), p. 11. In light of the principle "as open as possible, as closed as necessary", the use of licenses to restrict possible uses seems in line with the obligations set by the Open Data directive. Cf, also, art. 8 ODD (n 121). conditions are objective, proportionate, non-discriminatory, justified on grounds of a public interest objective and not unnecessarily restrictive. On the contrary, by stipulating that "legitimate commercial interests, knowledge transfer activities and pre-existing intellectual property rights shall be taken into account", the provision suggests that a variety of licensing types can be used depending on the particulars of a dataset.

[141] Art. 28, § 3, g, GDPR

the GDPR with respect to the movement of data and, in particular, the rules on data transfers to third countries. Controllers are required to keep track of the recipients to which they transfer data, no matter whether the transfer is direct or occurs by making data publicly available. Moreover, in case in which the recipients are based in a third country, the lawfulness of the transfer depends on the existence of the grounds set by Chapter V of the GDPR. This means that controllers must have in place technical and organisational measures to assess whether data recipients are from third countries and, in such case, whether the transfer to such country is covered by an adequacy decision by the Commission, or alternatively, whether the transfer can be legitimised by one of the other grounds provided by Chapter V of the GDPR. Clearly, beside the need to respect the provisions on the movement of data, the necessity to "close the access" to data depends on the risks posed by the dataset, the other safeguards implemented, harmful downstream use, having restricted the possibility to further disseminate and re-use, an assessment of the effectiveness of the license, etc.. In this perspective, the adoption of measures to ensure the traceability of dataset respond to the need to satisfy the "data movement" requirements and the broader risk-approach set by the GDPR. The capacity to trace the movement of datasets simultaneously contributes to the compliance with the obligations set by the Open Data directive. The traceability measures imposed by the GDPR substantively overlap with the measures required to make data "Findable" under the FAIR principles[142]. In this sense, multiple sources stress the importance of using permanent unique identifiers of datasets, such as DOI[143]. Next to findability, permanent identifiers could play a further function in a data protection perspective[144]. The identifier could be provided to data subjects pursuant to art. 13 and 14 as a means to trace the movement of their data. Through licenses, the controller could require recipients to use the same identifier in their derivative datasets. This would also facilitate controllers' compliance with art. 19 GDPR.

In essence, the implementation of access management and traceability measures is the *conditio sine qua non* for controllers to become aware of and govern the downstream lifecycle of the datasets they release. It is worth highlighting that under the regime of responsibility established by the GDPR, controllers are required to perform a dynamic assessment of the risks posed by the processing. Accordingly, controllers are under an obligation to revise and update the technical and organisational measures aimed at mitigating risks. The by design approach enshrined by the GDPR requires controllers to be prepared and reactive in the cases in which "things go wrong". A stress test for the adequacy of the safeguards adopted is represented by controllers' self-assessment of their ability to effectively implement measures of dataset retraction or deprecation. Not only dataset retraction is in itself essential to limit the potential harms caused by datasets. Retraction-deprecation measures provide a further example of "requirement backpropagation". In order to ensure the effectiveness of future retraction measures, controllers must identify and implement ex ante the necessary design choices.

Recent studies on the rise and fall of popular datasets[145] have emphasised how the data science community lacks of common and effective practices of dataset retraction[146]. The scarce effectiveness of retraction measures is due especially to lack of structural safeguards capable of avoiding the uncontrolled dissemination, re-use and extension of datasets. Where retraction measures are unable

---

[142] "F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource". Wilkinson and others (n 77).

[143] EU Grants: Data Management Template (HE):V1.0 – 05.05.2021, https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/reference-documents;programCode=HORIZON, § 2.1.; Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information, C/2018/2375, http://data.europa.eu/eli/reco/2018/790/oj "§ 4. … ensuring that datasets are easily identifiable through persistent identifiers and can be linked to other datasets and publications through appropriate mechanisms, and that additional information is provided to enable their proper evaluation and use; … § 5 … unique identification (interlinking of research outputs, researchers, their affiliations and funders, and contributors) is promoted through a wide range of persistent identifiers, in order to enable findability, reproducibility and long-term preservation of the research results". See, also, Alexandra Sasha Luccioni and others, 'A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication', *2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) <https://doi.org/10.1145/3531146.3533086> accessed 28 April 2023. https://doi.org/10.1145/3531146.3533086. On the use of DOI, see Benjelloun et al (n 18)

[144] In this sense, cf the approach endorsed by the Commission in the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A European strategy for data, COM(2020) 66 final, Brussels, 19.2.2020, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52020DC0066, p. 10

[145] *Supra* (nn 14, 17, 20)

[146] Luccioni and others (n 139); Peng, Mathur and Narayanan (n 138).

to “follow” runaway data, copies of retracted datasets or of their derivative continue to be available[147] and used, also in good faith, by the community[148]. In order to ensure the effectiveness of retractions, it is necessary that controllers provide clear and specific information to dataset recipients. In this sense, Luccioni et al. propose the adoption off “dataset deprecation reports”[149] detailing the reasons of the retraction and clear grounded instructions as to the prohibition of re-use of the retracted dataset[150]. For such ex post measures to be successful, it is essential that the ex ante safeguards discussed above are in place. In the absence of access and traceability measures, taking down a dataset and reaching its recipients would be almost impossible. The putting in place of a by design approach is necessary to implement a retraction plan capable of preventing downstream use of the dataset and of the derivative dataset that are impacted by the retraction of the former[151].

[147] Luccioni and others (n 143).Peng, Mathur and Narayanan (n 138).
[148] Luccioni and others (n 139), p. 205; Peng, Mathur and Narayanan (n 134), p. 7
[149] Luccioni and others (n 139), p. 205
[150] Peng, Mathur and Narayanan (n 134), p. 5; Luccioni and others (n 143)., pp. 199-200
[151] Luccioni and others (n 143).; p. 200

3. Conclusions

The Report has argued that the level of legal protection enjoyed by the legal subjects located downstream ML-pipelines depends on the extent to which the practices of dataset creation, curation and dissemination incorporate by design safeguards. In the first Section we have illustrated how the incentive structure that characterises current ML practices can lead to the accumulation of a "technical debt" along the ML-pipeline. We have argued that such debt can eventually result in the development and deployment of data-driven systems that lack the safeguards necessary to avoid downstream harm to natural persons. In Section 2 we have illustrated how a non-siloed, forward-looking approach to compliance with data protection law can address the accumulation of Legal Protection Debt. We have claimed that the fundamental provisions of the GDPR establish a framework of liability under which controllers are required to constantly engage in an effort of anticipation and mitigation of the potential harms caused by data processing activities. We have pointed out that the intensity of the anticipation and mitigation effort required to controllers depends on the features and the context of the processing. In particular, we have stressed that controllers' decision-making process must take into account all the available scientific and technical knowledge. Such factor contributes to the determination of both i) the scope of the potential harms that controllers can be expected to foresee and ii) the adequacy of the measures adopted to address such harms. In Section 2.2. and 2.3. we have analysed some of the requirements established by the GDPR. We have illustrated how a holistic approach to the compliance with such requirements can facilitate the effective anticipation and mitigation of potential downstream harm. In particular, we have underlined the important role played by documentation practices. In this sense, we have illustrated how the GDPR requires controllers to keep documentation concerning, *inter alia*:

- the subjects actively involved in the processing
- the categories of data and categories of data subjects
- technical aspects of processing
- the purpose and legal basis for each distinct processing activity
- the storage period
- automated decision making
- the security measures adopted
- the measures adopted to comply with the provisions on data subjects' rights
- data transfers and the recipients of data
- risk assessment and mitigation measures
- the applicability of the special regime for processing performed for scientific research purposes

We have argued that such measures assume particular relevance in the perspective of the dissemination and potential re-use of research data. In this respect, we have highlighted how the adoption of the safeguards and documentation practices required by the GDPR assume relevance also under the Open science and Open data framework. We have claimed that the development of data dissemination and re-use practices capable of addressing the accumulation of Legal Protection Debt requires the combination of documentation measures and further safeguards. Among such safeguards, we have highlighted, in particular:

- the use of licenses
- the selection of adequate dataset repositories
- the implementation of access management and traceability measures
- the use of unique identifiers
- the preparation of adequate documentation for released dataset
- the planning of dataset retraction strategies

ANNEX: Processing personal data for scientific research purposes under the GDPR

In this Annex, we analyse the special regime set by the GDPR for the processing of personal data for scientific research purpose.
First, we briefly identify the rationale and the scope of application of the special regime.
Secondly, we illustrate the structure of the special regime, distinguishing three levels: i) the general provision contained in art. 89, § 1, GDPR; ii) a set of provisions setting derogations from GDPR requirements; iii) the further derogations which can be adopted under EU or Member States law.
Thirdly, we examine the relation between the special regime and the other provisions of the GDPR that are not derogated because of the scientific research purpose of the processing.

A.1. The rationale and scope of the special regime

Research is a freedom protected both at the EU[152] and Member States level[153]. One of the objectives of the European Union is that of achieving a European Research Area[154]. The value of research is incorporated into the GDPR through the establishment of a particular regime of derogations from the general disciplines that can find application in the cases in which personal data are processed for scientific research purposes. Through such special regime, the GDPR aims at adapting data protection requirements in order to meet the public interest served by research[155]. The special regime thus encourages research activities while ensuring that the safeguards established by data protection law are respected in line with the principles of necessity and proportionality. As the European Data Protection Supervisor puts it, "performing an activity deemed to be research cannot be a carte blanche to take irresponsible risks"[156].
Accordingly, the notion of scientific research under data protection law is broad but not unlimited. Recital 159 of the GDPR states that "the processing of personal data for scientific research purposes should be interpreted in a broad manner"[157]. A non-exhaustive lists of forms of research included in the scope of special regime includes:

- technological development and demonstration,
- fundamental research,
- applied research,
- privately funded research
- studies conducted in the public interest in the area of public health[158].

As it emerges from the examples, the notion of scientific research encompasses both publicly and privately funded research. Under the GDPR, scientific research can be performed by both no-profit and for-profit organisations[159]. At the same time, some requirements have been indicated by the EDPB and EDPS. As the European Data Protection Board (EDPB) points out, scientific research in the context of data protection law is to be understood as "a research project set up in accordance with

[152] Charter of Fundamental Rights of the European Union, http://data.europa.eu/eli/treaty/char_2012/oj, art. 13: "The arts and scientific research shall be free of constraint. Academic freedom shall be respected."
[153] For an overview of the constitutional provision on the freedom of research in Member States: https://fra.europa.eu/en/eu-charter/article/13-freedom-arts-and-sciences#national-constitutional-law
[154] Consolidated version of the Treaty on the Functioning of the European Union, http://data.europa.eu/eli/treaty/tfeu_2016/oj, art. 179, § 1.
[155] European Data Protection Supervisor (EDPS), A Preliminary Opinion on data protection and scientific research, 6 January 2020, https://edps.europa.eu/sites/default/files/publication/20-01-06_opinion_research_en.pdf
[156] Ivi, p.12
[157] GDPR, Recital 159
[158] Ivi
[159] Cf also Giovanni Comandè and Giulia Schneider, 'Differential Data Protection Regimes in Data-Driven Research: Why the GDPR Is More Research-Friendly Than You Think' (2022) 23 German Law Journal 559.

relevant sector-related methodological and ethical standards, in conformity with good practice"[160]. In a similar perspective, the EDPS considers that scientific research is "the research … carried out with the aim of growing society's collective knowledge and wellbeing, as opposed to serving primarily one or several private interests"[161].
The rationale of the special regime is strictly tied to the instrumental role that research plays for the achievement of public goals. Accordingly, such instrumentality also delimits the scope of the special regime. The derogations provided by the special regime "are strictly connected to research purposes [and therefore] cannot spill over other data processing purposes"[162]. The principle of segregation[163] established by art. 89, § 4, GDPR implies that, if the controller processes data for a different, e.g., commercial, purpose, the ordinary requirements established by the GDPR expand back.

A.2. The structure of the special regime

The special research regime established by the GDPR results from the effects of multiple provisions. Such provisions do not apply in block and the special regime is shaped according to a multi-layered structure articulated as follows:

a) the general norms established by art. 89, § 1, GDPR apply to all processing that are carried out for scientific research purposes;
b) several provisions of the GDPR contain, next to a general discipline, a set of specific provision that apply only in cases of processing for scientific research purposes;
c) provisions that empower the EU and national lawmaker to establish further derogations, i.e.: i) the general provision established by art. 89, § 2, GDPR; ii) the exception to the prohibition of processing of particular categories of data provided by art. 9, § 2, j, GDPR.

A.2.1. The first layer: the general norm provided by art. 89, § 1, GDPR

The cornerstone of the special research regime is represented by art. 89, § 1, GDPR. This provision establishes a set of requirements that apply to any processing of personal data for scientific research purposes. It is important to highlight that compliance with the obligations established by such provision is relevant in itself and also as a requirement necessary to enjoy the derogations set by all the other provisions that form the special regime. Art. 89, § 1, is referred to by all the provisions of the special regime and the derogations thereby provided can find application only "subject to the conditions and safeguards referred to in", or "in accordance with", or "pursuant to", Article 89, § 1.
The first part of art. 89, § 1, requires controllers processing personal data for scientific purposes to adopt appropriate safeguards for the rights and freedoms of data subjects. This is a wide-ranging, context-dependent, obligation that informs all the relevant organisational and technical decisions that controllers are called to make in designing their processing activities. Art. 89, § 1, GDPR specifies that the safeguards that controllers are required to adopt must include technical and organisational measures capable of ensuring, in particular, the respect of the principle of data minimisation. While art. 89, § 1, GDPR does not provide a list of measures, the text of the provision indicates that such measures may include pseudonymisation and anonymisation. By providing the example of pseudonymisation and anonymisation, art. 89, § 1, offers a set of criteria that can guide the choice of further measures. In this sense, the decisions as to the safeguards to be implemented must be based on the consideration of the specific purpose of the research. On this basis, controllers are required to perform a necessity test aimed at establishing the extent to which the processing of *personal* data is necessary to achieve the research purpose. Accordingly, controllers should first establish whether their purpose can be fulfilled through the processing of anonymous or anonymised data and, in such case, avoid the processing of personal data. Otherwise, controllers should assess whether the research purposes can be fulfilled through pseudonymised data. In essence, coherently with the core tenets of data protection law, the processing of personal data should be allowed only where

[160] European Data Protection Board, Guidelines 05/2020 on consent under Regulation 2016/679, Version 1.1, Adopted on 4 May 2020, https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf, p. 30, § 153. See, also Giovanni Buttarelli (EDPS), Fifth World Congress for Freedom of Scientific research, 12 April 2018, https://freedomofresearch.org/wp-content/uploads/2018/04/Giovanni-Buttarelli-Speech.pdf
[161] EDPS (n 155), p. 12
[162] Comandè and Schneider (n 159).
[163] ibid.

necessary and proportionate. In this sense, art. 89 does not introduce a new requirement with respect to those that govern any processing. The emphasis put by art. 89 on anonymisation and pseudonymisation, however, seems to require controllers to comply to a stricter standard with respect to processing performed for other purposes. Such a stricter standard is justified in light of the derogations that we will analyse in the following paragraphs.

It is worth repeating that, as discussed *supra*, pseudonymization and anonymisation are themselves processing of personal data. Pseudonymization and anonymisation are a peculiar form of processing in that their successful performance can result in the application of a different legal regime to the input and output data of such processing[164]. Successful anonymisation results in non-personal data and, therefore, processing of the latter is not governed by data protection law. As discussed above[165], the definition of personal data is such that achieving anonymisation is particularly difficult. It is more likely that controllers can achieve (strongly) pseudonymised data. As discussed, pseudonymised data are still personal data. However, controllers processing pseudonymised data can enjoy a lessening of some of the obligations established by the GDPR[166]. It is important to highlight that such reduction of the compliance burden of controllers only applies to the pseudonymised data, that is, the data that "can no longer be attributed to a specific data subject without the use of additional information"[167]. For the latter to be considered pseudonymised, the additional information must be kept separately and be subject to technical and organisational measures that prevent identification[168]. The GDPR applies in full to such separate information as long as it constitutes in itself personal, non-pseudonymised data. At the same time, the GDPR applies in full to the processing performed *before* that the result pseudonymisation is achieved. Art. 11 GDPR is particularly relevant in cases of processing of personal data that do not allow the identification of data subjects, as pseudonymised data. Where the controller processes pseudonymised data for a purpose that does not or no longer requires the identification of data subjects, such controller is relieved from the obligation to identify the data subject by maintaining, acquiring or processing additional information for the sole purpose of complying with the GDPR. If the controller can demonstrate that it cannot identify data subjects, the controller can also be lifted from the obligations provided in articles 15 to 20[169], unless the information necessary for the identification is provided by a data subject for the purpose of exercising his or her rights under those articles.

As said, while art. 89, § 1, expressly refers only to pseudonymisation and anonymisations, these are not the only measures which controllers may be required to adopt in order to comply with art. 89, § 1. In line with the general norm established by art. 24 GDPR[170], controllers enjoy a large discretion – and a corresponding responsibility – with respect to the choice of the means to ensure compliance with the obligation to put in place safeguards for the rights and freedoms of the data subject and technical and organizational measures to ensure the principle of data minimization and the other requirements of data protection law. In this sense, we maintain that compliance with the requirements established by art. 89, § 1, is governed by the same framework that we have examined in section 2. Accordingly, we refer the readers to the recommendations formulated in the Report.

A.2.2. The second layer: derogations provided by specific provisions of the GDPR

As anticipated, the second layer of the special regime for research is regulated by a set of provisions of the GDPR. Such provisions establish a general discipline and a set of requirements the satisfaction of which allows the application of a derogatory regime. Such derogatory provisions concern:

- the principle of storage limitation (art. 5, § 1, e);
- the principle of purpose limitation (art. 5, § 1, b);
- the obligation to provide information where data are obtained from sources other than the data subjects (art. 14, § 5, d);
- the right to erasure (art. 17, § 3, d);

[164] See section 1.2.1., fig. 3
[165] Section 1.2.
[166] Ivi
[167] Art. 4, § 1, n. 5, GDPR
[168] Ivi
[169] Art. 11, § 2, GDPR
[170] Section 2

- the right to object (art. 21, § 6).

A.2.2.1. Derogation from storage limitation obligation

Art. 5, § 1, e, GDPR, requires data controllers to not keep personal data in a form that permits the identification of the data subject for longer than it is necessary for achieving the purpose of the processing. The same article provides that, "insofar as the personal data will be processed solely for … scientific … research purposes", personal data can be stored for longer periods. The applicability of such derogation is conditioned to the compliance with art. 89, § 1, i.e., "subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject".

A.2.2.2. Derogation from purpose limitation obligation. Further processing, curation and off-the-shelf datasets

Purpose limitation represents a pillar of data protection law, as provided by art. 8, § 2, of the Charter of Fundamental Rights of the European Union[171] and art. 5, § 1, b, GDPR. Purpose limitation implies two requirements: i) the collection of personal data is prohibited unless it is performed for purposes that are "specified", "explicit", and "legitimate" – purpose specification requirements; ii) personal data cannot be further processed in a manner that is incompatible with the specified", "explicit", and "legitimate" purposes of the collection - purpose compatibility requirement[172].

The purpose specification requirement means first of all that data protection law rules out the possibility that personal data are processed "without a purpose", i.e., a vague or undefined purpose is still a purpose. Data protection law demands controllers to make their best effort to specify such purpose[173]. At the same time, controllers' statements about the purpose of the processing do not have a constitutive effect: whatever controllers say about the purpose of the processing, the actual purpose of the processing can be inferred – e.g., by DPA or judicial authorities - from other elements.

A specified, explicit and legitimate purpose is an essential requirements for processing personal data. Controllers are required to informs data subjects of the purpose of the processing and of its corresponding legal basis at the moment of the collection or obtaining of the data[174]. Through the information provided by the controller, data subjects develop legitimate expectations as to the processing to which they are data will be subject: who will process their data, for what specified purpose, based on what legal basis. This leads us to the requirement of purpose compatibility. If the controller processes data for a purpose different and incompatible with that communicated to data subjects, the expectations of the latter would be frustrated. Accordingly, the GDPR prohibits the processing of personal data in a manner that is incompatible with the initial purpose for which data were collected or obtained.

The requirements established by the provision on purpose limitation are subject to derogations under the scientific research regime. First, the GDPR acknowledges that processing performed for scientific research purposes may justify a lessening of the purpose specification requirement. In particular, Recital 33 of the GDPR recognizes that, in case of scientific research, the full identification of the purpose of processing might often not be possible at the time of the collection of data. Accordingly, the Recital invites to allow data subjects to give their consent to "certain areas of research or parts of projects", under the condition of the respect of ethical standards for scientific research. This Recital addresses a specific issue emerging at the moment of the collection of personal data, with reference to the giving of consent. Therefore, the aim of the Recital should be understood as that of ensuring the validity of the consent given to processing for research purposes, and not as that of allowing a generalized restriction of the requirement of purpose specification. This means that, as soon as the purpose of the processing becomes more clear, such purpose has to be specified and communicated to the data subjects.

Secondly, the GDPR, establishes a presumption of purpose compatibility for the cases in which the purpose of the processing is scientific research. Art. 5, § 1, b, provides that

[171] The Charter of Fundamental Rights of the European Union (http://data.europa.eu/eli/treaty/char_2012/oj), art. 8, § 2, provides that "data must be processed fairly for specified purposes".

[172] Art. 5, § 1, b, GDPR; Article 29 Working Party, Opinion 03/2013 on purpose limitation, Adopted on 2 April 2013, 569/13/EN, WP 203, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[173] Cf also art. 12 GDPR, with respect to the obligation to inform data subjects of the purpose of the processing pursuant to art. 13, 14, 15 GDPR

[174] Art. 13, § 1, c, and 14, § 1, c, GDPR.

> further processing for … scientific … research purposes … shall, in accordance with Article 89, § 1, not be considered to be incompatible with the initial purposes[175]

As Manis highlights, "[t]he rationale behind this presumption is to provide researchers and research organizations with a certain freedom to re-examine dataset collected or processed in the context of a previous scientific project"[176].As the text of the provision makes clear, the application of the presumption is conditioned to the respect of some requirements that we will analyse shortly. At the same time, it is necessary to highlight that the presumption established by art. 5, § 1, b, is non-absolute, i.e., it can be rebutted whenever the specific circumstances of the processing are such as to make the purpose of the further processing incompatible with the purpose for which the data were collected or obtained.

To better clarify the operation of the presumption of compatibility it is opportune to briefly discuss the general discipline established by the GDPR for the cases of further processing[177]. Under the GDPR, the processing of personal data for purposes other than those for which the personal data were initially collected or obtained should be allowed only where the former purpose is compatible with the initial purposes[178]. Accordingly, the controller who intends to perform further processing is required to carry out a purpose compatibility test[179]. First, the controller has to make sure that the original processing meets all the requirements that ensure the lawfulness of the latter[180]. Then, the controller shall perform a comprehensive assessment of the purpose of the further processing in light of the purpose of the initial processing. Controllers are required to document both the assesment made and the decision adopted[181]. Art. 6, § 4, and Recital 50 provide a non-exhaustive list of factors that the controller should take into account in performing the compatibility test, i.e.:

(a) the existence of any link between the purposes for which the personal data have been collected and the new intended purposes;
(b) the context of the data collection and, in particular, the reasonable expectations of data subjects based on their relationship with the controller as to their further use[182];
(c) the nature of the personal data;
(d) the possible consequences of the intended further processing for data subjects;
(e) the existence of appropriate safeguards, which may include encryption or pseudonymisation, in both the original and intended further processing operations[183].

The possible outcomes of the compatibility test are the following:

- **Non compatibility:** where the purpose of the intended processing is incompatible with the purposes for which data were collected, the controller needs a new legal basis to further process the data. In this case, additional caution is required. The controller must carefully consider the legal basis relied upon for the initial processing. Where the only legal basis grounding the initial processing was the consent of the data subject, the controller cannot swap to a different basis, e.g. legitimate interest[184]. This would frustrate the legitimate expectations of the data subject. Therefore, in such cases, the controller is required to obtain a new consent from the data subject.

---

[175] Recital 50 GDPR: "…Further processing for … scientific … research purposes … should be considered to be compatible lawful processing operations".
[176] Maria Luisa Manis, 'The processing of personal data in the context of scientific research. The new regime under the EU-GDPR' [2017] BioLaw Journal - Rivista di BioDiritto 325. https://doi.org/10.15168/2284-4503-259, at 331
[177] Art. 5, § 1, b, and art. 6, § 4, GDPR.
[178] Recital 50 GDPR
[179] Art. 6, § 4, GDPR. According to the same provision, the controller can perform further processing for a purpose other than that for which the personal data have been collected without carrying out a compatibility test where: i) data subjects consent to the further processing; or ii) the further processing is provided by "Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in art. 23, § 1". In such cases the further processing satisfies the requirements of purpose and legal basis due to either the new consent or the normative provision.
[180] Recital 50 GDPR
[181] Section 2.2
[182] Recital 50 GDPR
[183] Ivi
[184] On the relation between consent and other legal bases, see European Data Protection Board, Guidelines 05/2020 on consent under Regulation 2016/679, Version 1.1, Adopted on 4 May 2020, https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf, p. 25: § 122-123: "It is important to note here that if a controller chooses to rely on consent for any part of the processing, they must be prepared to respect that choice and stop that part of the processing if an individual withdraws consent. Sending out the message that data will be processed on the basis of consent, while actually some other lawful basis is relied on, would be fundamentally unfair to individuals. In other words, the controller cannot swap from consent to other lawful bases. For example, it is not allowed to

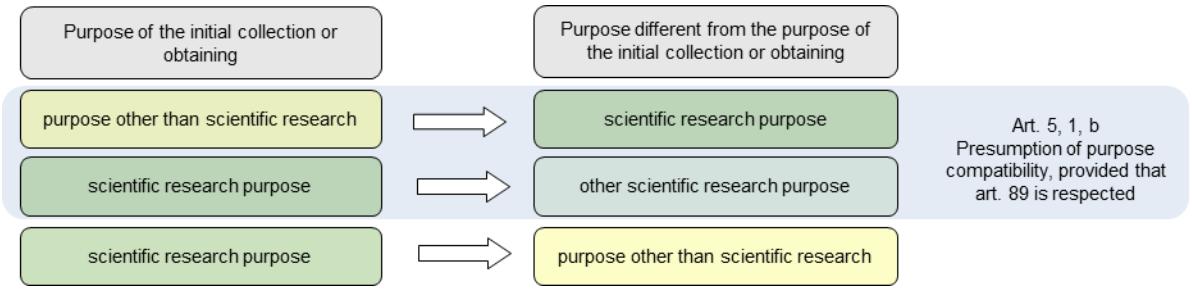
● **Compatibility**: where the purpose of the intended processing is compatible with the purposes for which data were collected, the further processing can be performed without the need of a different legal basis separate from that which allowed the collection of the personal data[185]. The further processing must comply with all the requirements set by the GDPR. In particular, the data controller must comply the information obligations, i.e., inform data subjects, prior to proceeding to the further processing, of the purposes of the new processing and of the rights of data subjects with respect to such processing[186].

As anticipated, the compatibility of the further purpose of processing is presumed where the purpose of the further processing is scientific research. However, the applicability of the presumption is dependent on the satisfaction of further requirements:

1) **Compliance with the obligations set by Article 89, § 1, GDPR:** the presumption of compatibility applies only where the controller has adopted appropriate technical and organisational safeguards, such as pseudonymisation and access limitations[187]. The Article 29 Working Party has also stressed the need that controllers ensure that data processed will not be used to support measures or decisions regarding any particular individuals[188];

2**) Lawfulness of processing**: "purpose specification and lawfulness of processing are separate and cumulative requirements"[189]. This means that the further processing of data, even where supported by the presumption of compatibility of scientific research purposes, requires a specific lawful ground[190]. Accordingly, controllers have to make sure that the intended further processing can be grounded on the same legal basis of the first processing.

Figure A.1.:Further processing by the same controller



The requirements descending from the principle of purpose limitation apply in a controller-focused perspective[191] to both the cases of self-collection of data, i.e., where data are collected directly from

retrospectively utilise the legitimate interest basis in order to justify processing, where problems have been encountered with the validity of consent. Because of the requirement to disclose the lawful basis, which the controller is relying upon at the time of collection of personal data, controllers must have decided in advance of collection what the applicable lawful basis is.

[185] Recital 50 GDPR

[186] GDPR, art. 13 and 14; Recital 50; European Data Protection Supervisor, A Preliminary Opinion on data protection and scientific research, 6 January 2020, https://edps.europa.eu/sites/default/files/publication/20-01-06_opinion_research_en.pdf , p. 20

[187] Section A.2.1; 2.3.

[188] Article 29 Working Party, Opinion 03/2013 on purpose limitation, Adopted on 2 April 2013. 13/EN, WP 203, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf, p.28

[189] European Data Protection Supervisor, A Preliminary Opinion on data protection and scientific research, 6 January 2020, https://edps.europa.eu/sites/default/files/publication/20-01-06_opinion_research_en.pdf , p. 22

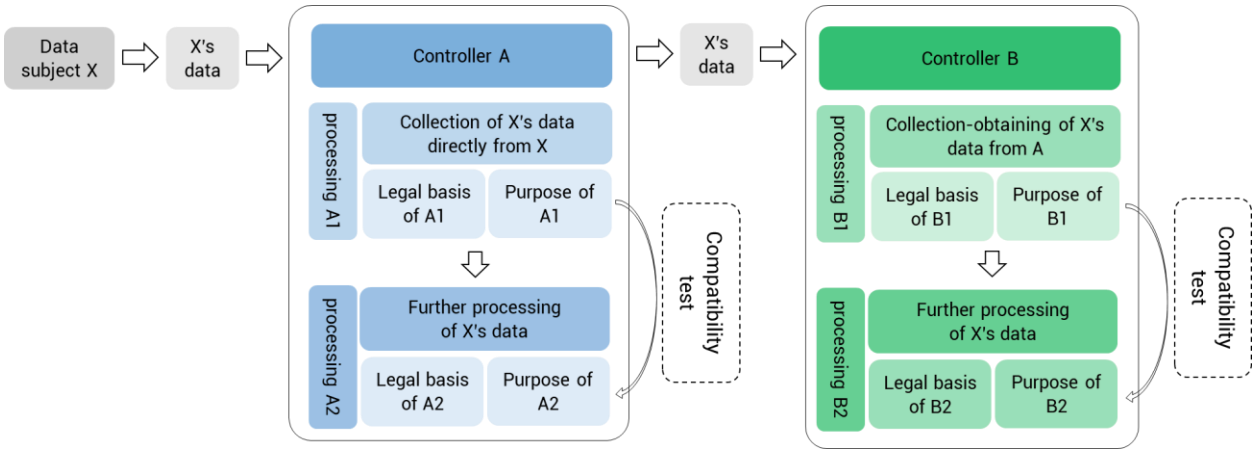
[190] Ivi

[191] The meaning of the phrasing "further processing" (art. 5, § 1, b) and "processing for a purpose other than that for which the personal data have been collected" (art. 6, § 4) have given rise to doubts. A systematic interpretation of the text of the GDPR leads to the conclusion that both expressions should be read as being followed by the words "by the same controller". This interpretation is supported by Recital 50, that states that, in case of purpose compatibility, "the further processing can be performed without the need of a different legal basis separate from that which allowed the collection of the personal data". The meaning of such Recital can be given a coherent interpretation only where understood as referring to cases in which the data collection has been performed by the same controller who now intends to further process the data. Such controller can indeed already rely on a legal basis for the collection-processing. Our interpretation seems to find support in Regina Becker and others, 'Secondary Use of Personal Health Data: When Is It "Further Processing" Under the GDPR, and What Are the Implications for Data Controllers?' (2022) 1 European Journal of Health Law 1. DOI:10.1163/15718093-bja10094, at 10: "In the controller-focussed view, each phase within the data lifecycle begins when a controller collects personal data — either directly

the data subjects, and the cases of obtaining of data, i.e., where data are collected from sources different than the data subject. Accordingly, also the presumption of compatibility applies in a controller-focused perspective. This means that, in order to lawfully obtain - and further process - data from sources other than data subjects, controllers must rely upon a legitimate purpose and legal basis, even where the purpose is scientific research. Where a processing is performed by a controller different than the controller who has initially collected the data, such processing cannot be considered as a "further processing", but as another processing *in toto*. This means first of all that the new controller cannot rely on the legal basis used by the previous controller to ground her processing. The obtaining of the data establishes a new, autonomous, relation between the new controller and the data subject. This new relation subject-controller is subject to all the requirements of GDPR examined supra, e.g., legal basis, purpose, transparency, etc.. In particular, the controller is required to ground the processing on a legal basis autonomous from the legal basis which grounded the lawfulness of the processing carried out by the previous controller. The legal bas(es) on which the old controller has grounded the collection, processing, and making available of the data contained in a dataset does not extend to the processing carried out by the new controller. Legal basis and purpose are inseparably tied to a *specific processing* carried out by a *specific controller*. The fact that the old controller was lawfully processing a subject's data based on her consent or other legal basis for a certain specified purpose does not make the processing carried out by the new controller lawful. Neither an assessment of compatibility of the purpose nor the circumstance that such purpose is scientific research can provide the new controller with the legitimate grounds for processing personal data of another controller.

The effects of the presumption of purpose compatibility for further processing can be enjoyed only by the same controller that has collected the data – either through self-collection or by obtaining them from another controller. In case of obtaining of data, the provisions on further processing and purpose compatibility find application only with respect to the relation between: i) the purpose for which the data were initially *obtained* and ii) the different purpose for which the controller intends to process the *obtained* data.

Figure A.2.: Scope of purpose compatibility test and compatibility presumption



The controller-focused scope of the provisions on purpose compatibility does not imply a reduction of the level of protection that both the old controller that transfers data and the new controller that obtains the data are required to ensure.

---

from the data subject or, in the case of existing data, from another source — and ends with the realisation of the purpose(s) for which that controller collected the data…the GDPR does not look onto the entire lifecycle but only onto stages in that lifecycle where the processing is determined by a single or by joint controllers. The GDPR focusses always on the fact that data are processed and why they are processed, building its definitions and framework around the processing operations and the purposes. … In line with this approach, further processing under the GDPR is to be understood in relation to the purpose for which a particular controller originally collected the data, whether directly from the data subject or by obtaining existing data from another source".

For the first controller, the norms on purpose compatibility apply to the further processing consisting in the transfer of data. Where controllers intend to transfer data availing themselves of the presumption of compatibility of scientific research purposes, such controllers must put in place the measures necessary to ensure that the transfer will actually be for scientific research purposes[192].
For the new controller obtaining the data, the decisions to be taken as to the legal basis necessary to perform the intended processing absorb most of the factors to be assessed in performing the compatibility test[193]. This applies, in particular, where the legal basis that the new controller is considering is legitimate interest[194]. The applicability of such legal basis requires the performance of a test aimed at assessing whether the legitimate interest of the (in this case, new) controller is overridden by the interest or fundamental rights and freedoms of the data subject. In performing the test, consideration should be given not only to the worthiness of the processing intended by the new controller - as it were, the interest has to be *legitimate*. As Recital 47 makes clear, controllers "to which the personal data may be disclosed, or …a third party" must consider "the reasonable expectations of data subjects based on their relationship with the controller". In particular, pursuant to Recital 47, controllers must assess

> "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. The interests and fundamental rights of the data subject could in particular override the interest of the data controller where personal data are processed in circumstances where data subjects do not reasonably expect further processing"[195].

Due to the indirect and less proximate relation controller-data subject, the obtaining of data from sources different than the data subject risks putting under stress the protection of the latter. This circumstance calls for the adoption of mitigating safeguards. This is particularly the case where further derogations established for processing performed for research purposes find application. For instance, controllers' decisions concerning legal basis and further processing should be particularly cautious where data subjects have not been directly provided with information on the processing pursuant to the derogation provided by art. 14, 5, b. We will examine such derogation in the next paragraph.

A.2.2.3 Derogations to data subject rights provided by the GDPR

In the following paragraphs, we analyse three provisions of the GDPR that, in case of processing for scientific research purposes and provided the respect of certain conditions, limit data controllers' obligations with respect to some of data subjects' rights.

A.2.2.3.1. The derogation to information obligations. The case of obtaining off-the-shelf datasets

A first derogation to data subjects' rights concerns controllers' obligation to provide data subjects with information concerning the processing. The transparency of the processing in relation to the data subject represents one of the core requirements that governs data protection law[196]. The provision of transparent information to data subjects is a pre-requisite for the lawfulness of processing[197], and also the content of specific information obligations[198].
Art. 14 GDPR disciplines the information obligations of controllers in the cases in which they process personal data obtained from sources different than the data subjects, thus also the case of obtaining of off-the-shelf datasets. As discussed in § A.2.2.2, if a dataset contains personal data, obtaining the dataset already constitutes "processing" under the GDPR, thereby triggering the application of data protection law. The new controller/recipient is subject to a duty to provide data subjects with information concerning the processing. In many respects, such duty mirrors the duty to provide information to data subject in the case in which the data are collected directly from the data subject. In particular, the controller is bound to provide data subjects with information concerning, *inter alia*, the

[192] Section 2.3.
[193] *supra*
[194] Art. 6, § 1, f, GDPR
[195] Recital 47, GDPR
[196] Art. 5, § 1, a, GDPR
[197] Especially where the legal basis of the processing is consent, pursuant to art. 6, § 1, a, and 7, GDPR, and where consent is, pursuant to art. 9, 2, a, the ground used to lift the prohibition of processing of special categories of data
[198] Art. 12; art. 13-15, GDPR

purposes and the legal basis of the processing[199]. Additionally, controllers processing obtained data are required to inform data subjects of the sources from which the personal data originate, and if applicable, whether the data have been obtained from publicly accessible sources[200]. In case of obtained data, the fulfilment of the information obligations established by art. 14 GDPR is particularly important. In these cases, a certain distance separate controllers and data subjects: the latter might not have had any contact with the former. The fulfilment of information obligation and, in particular, the disclosure of the origins of the data, provides data subjects with means necessary to monitor the processing and circulation of their data. This, in turn, allow data subjects to demand the respect of their rights against both the controller *a quo* and the controller *ad quem*.

This being the general discipline, art. 14, § 5, GDPR introduces a set of circumstances under which the controller obtaining data from sources other than the data subjects might be relieved of the obligation to *provide* information to the data subjects. The italic added is intended to emphasise that, as we will illustrate shortly, the only obligation that the new controller can derogate is that of *providing* information, not of *having* such information[201]. Among the circumstances listed by art. 14, § 5, a specific derogations is established for the cases in which the processing of obtained data is performed for scientific research purposes. Pursuant to art. 14, § 5, e, the exception can find application where a set of requirements is cumulatively met. As said, the first requirement is that the purpose of the processing is scientific research. It is worth emphasising that, pursuant to the principle of segregation governing the special research regime, this conditions is met only to the extent that scientific research is the exclusive purpose of the processing.

The second requirement for the applicability of the derogation is that controllers prove the existence of one of the following alternative circumstances:

- providing the information to the data subjects proves impossible, or
- providing the information to the data subjects would involve a disproportionate effort, or
- the obligation referred to in art. 14, § 1, is likely to render impossible or seriously impair the achievement of the objectives of the processing.

Since that provided by art. 14, § 5, is a special regime that derogates provisions implementing the principle of transparency, the scope of such provision should be interpreted strictly[202]. First, the derogation is, as usual, conditioned to the respect of the requirements set by art. 89, § 1. The exemption provided by art. 14, § 5, GDPR does not provide a carte blanche to controllers: its rationale is that of facilitating the achievement of research purposes where the research requires the processing of data of a high number of subjects and such data have been subject to high measures of pseudonymisation[203]. Art. 14, § 5, e assumes particular relevance in the cases in which the combined effect of art. 4, § 1, n. 5 and art. 11 GDPR, that is, cases in which the new controller obtains pseudonymised data, i.e., data that can no longer be attributed to a specific data subject without the use of additional information that available only to the old controller[204]. Where the processing carried out by the new controller does not require the identification of data subjects, such controller is not obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with the GDPR[205]. Secondly, the actual existence and relevance of the circumstances listed in art, 14, § 5, e, must be proved by the controller, in accordance to the principle of accountability set by art. 24 GDPR. In particular, the fact that art. 14 refers to the circumstance in which providing the information to the data subjects "*proves impossible*" means that an attempt in such sense must have been done. Equally, controllers must prove the "disproportionate character" of the effort required, and the likeliness that providing the information to the data subject would "render

[199] Art. 14, § 1, c, d, GDPR

[200] Art. 1, § 2, f, GDPR

[201] Moreover, controllers are not exempted from the obligation to eventually disclose such information, e.g., in the case of a request by the DPA or in case of an access request by the data subject pursuant to art. 15 GDPR. Even if art. 89(2) provides that in case of processing for scientific research purposes EU or Member State law may provide for derogations to, inter alia, the right to access, such derogations are, once again, legitimate only in so far as such right is "likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes". This means that the fact that a controller is processing personal data for scientific research purposes does not automatically implies the applicability of the derogation.

[202] European Data Protection Board, Guidelines on transparency under Regulation 2016/679, Adopted on 29 November 2017, As last Revised and Adopted on 11 April 2018, 17/EN, WP260rev.01, https://ec.europa.eu/newsroom/article29/redirection/document/51025, pp. 28-31

[203] Paul Quinn and Liam Quinn, 'Big Genetic Data and Its Big Data Protection Challenges' (2018) 34 Computer Law & Security Review 1000, at 1014 .

[204] i.e., "are kept separately and subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person", GDPR, Recital 26

[205] Art. 11, § 1, GDPR

impossible or seriously impair the achievement of the objectives of that processing". This means both: i) that controllers must be able to demonstrate that they have made all reasonable efforts; ii) that whenever providing the information to data subject is possible, even if it requires *a certain effort*, it must be provided. The rationale of the derogation is that of sparing researchers from wasting of resources – e.g., renounce to a research project or start it over – where the availability of data is not paired with the possibility to contact data subjects. However, this does not mean that controllers can negligently give cause to the impossibility or difficulty to inform data subjects. That is, impossibility or difficulty cannot be the result of research design choices that could have been done differently.
In the case in which the abovementioned conditions are met, controllers can enjoy the lessening of the information obligations established in art. 14, §§ 1-4. The application of the derogatory regime, however, does not imply the exemption from *any information obligation*. Art. 14, § 5, b, provides that, when availing themselves of the derogatory regime, controllers shall nonetheless:

- ensure that "the conditions and safeguards referred to in Article 89, § 1" are in place;
- "take appropriate measures to protect the data subject's rights and freedoms and legitimate interests"

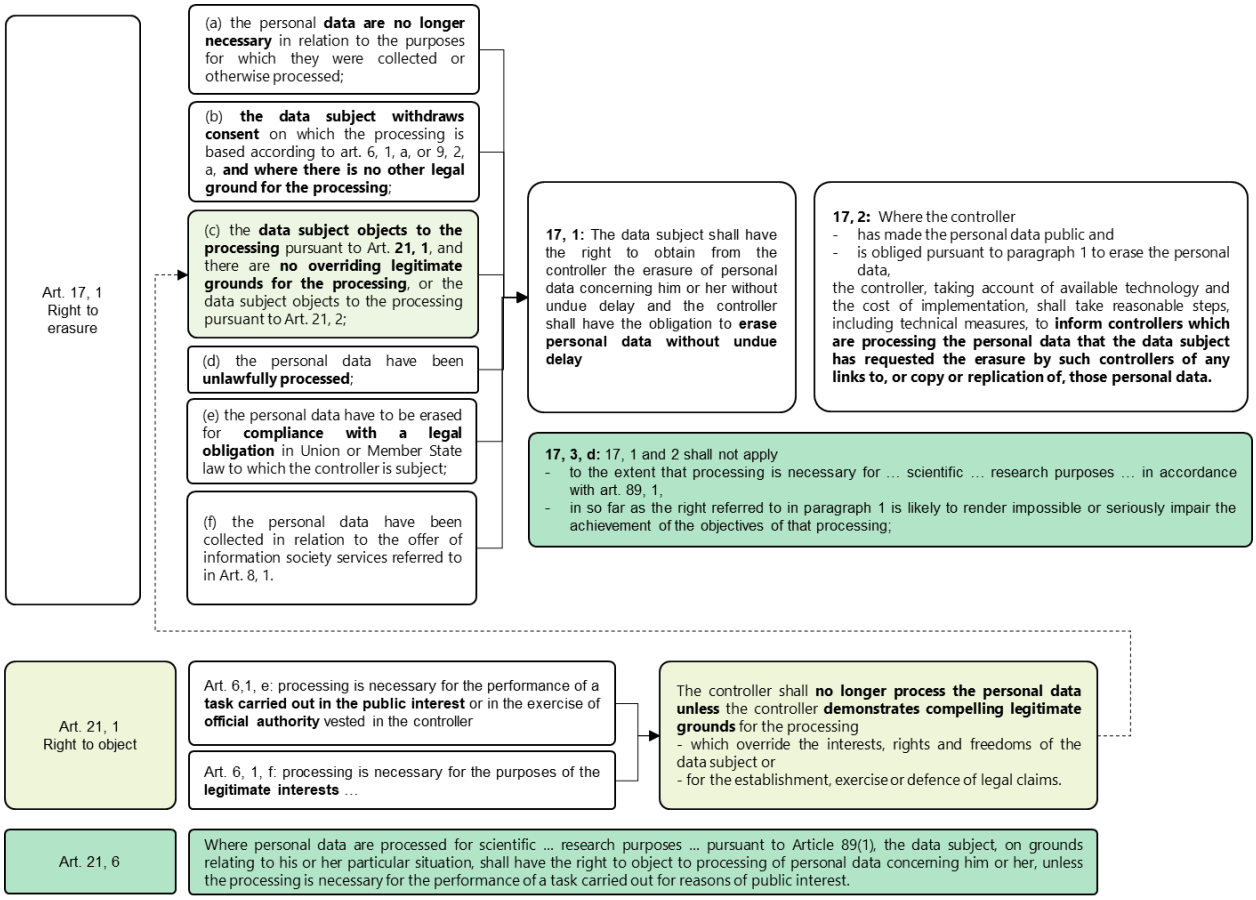
Among such "appropriate measures", art. 14, § 5, b, specifically mentions "making the information publicly available". This means, at least, the publication of such information on the website of the research project and/or the research institution. Other measures might include the documentation and traceability measures discussed above in the Report[206].
As already stressed, the only obligation that the new controller can be dispensed with pursuant to art. 14, § 5, is the obligation to *provide* the information *directly* to the data subjects. The new controller is in any case bound to be in possession of the information required by art. 14. Having such information is necessary to demonstrate compliance with under art. 24 GDPR and, if applicable, it must be kept in the record of processing activities pursuant to art. 30 GDPR[207]. Once again, the new controller must have determined, inter alia, the purpose of processing and must have made sure that the processing is grounded on one of the legal bases provided by art. 6 GDPR *before* obtaining the data[208].

A.2.2.3.2. Right to erasure (art 17, § 3) and right to object (art. 21, § 6)

Art. 17 and 21 GDPR establish an interlocking set of remedies aimed at ensuring data subjects' control on the processing. In a nutshell, depending on the legal basis of the processing, data subjects have different instruments to intervene and potentially terminate the processing. Where the processing is based on the consent of the data subject, the latter can withdraw the consent and obtain the erasure of her data, unless a different legal basis is available for the controller to continue the processing[209]. Where the processing is based on legitimate interest or, *latu sensu*, public interest, data subjects can exercise the right to object and, if successful, obtain the erasure of their data[210].

Figure A3: Art. 17 and 21 GDPR

[206] Cf also section 2.2 and 2.3.
[207] Cf section 2.2.
[208] Cf. European Data Protection Board, Guidelines 05/2020 on consent under Regulation 2016/679, Version 1.1, Adopted on 4 May 2020, https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf, p. 25, § 121 and fn 59, according to which the application of one of the six legal bases "must be established prior to the processing activity and in relation to a specific purpose. Pursuant to Articles 13 (1)(c) and/or 14(1)(c), the controller must inform the data subject thereof."
[209] Art. 17, § 1, b, GDPR
[210] In the cases in which the processing is based on lett. b, c, d, of art. 6, § 1, GDPR (i.e., respectively, processing necessary for the performance of a contract, processing necessary for compliance with a legal obligation, processing necessary in order to protect the vital interests of the data subject or of another natural person), an erasure request can be grounded on lett. a) and d) of art. 17, § 1, i.e., respectively, where the personal data are no longer necessary in relation to the purposes for which they were collected or otherwise processed; or where the personal data have been unlawfully processed.



The right to erasure established by art. 17 performs a function that is broader than that of providing a sanction to the withdrawal of consent to processing. In essence, the requirements provided by art. 17, § 1, let. a-f correspond to core requirements of data protection law. For instance, in line with the requirement of storage and purpose limitation, data subjects have a right to obtain the erasure of their data when data are "no longer necessary in relation to the purposes for which they were collected or otherwise processed"[211]. More in general, data subjects have a right to erasure when "personal data have been unlawfully processed"[212]. On a similar level of generality, data subjects have a right to obtain the erasure of their personal data when the latter have to be erased in order to comply with a legal obligation[213]. Art. 17, § 2, introduces further obligations for the cases in which controllers have made personal data public and, at a second time, they are obliged to erase such data pursuant to art. 17, § 1. Pursuant to art. 17, § 2, based on the consideration of the available technology and the cost of implementation, controllers are required "to take into account and to take all reasonable steps to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data".

As said, the entitlement to a right to data erasure can also be the legal consequence of the successful exercise of the right to object pursuant to Article 21[214]. Under such provision, data subjects can object to the processing of their data based on "grounds relating to [their] particular situation". This remedy is granted to data subjects in the cases in which the processing is grounded on the legal basis of the legitimate interests or when the processing is "necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller". When receiving data subjects' objection to the processing, controllers are required to assess whether: i) there exist "compelling legitimate grounds for the processing which override the interests, rights and freedoms of

[211] Art. 17, § 1, a, GDPR
[212] Art. 17, § 1, d, GDPR
[213] Art. 17, § 1, e, GDPR
[214] Art. 17, § 1, c, GDPR

the data subject"; or ii) the processing is necessary "for the establishment, exercise or defence of legal claims". Unless controllers can demonstrate the presence of one of such grounds, they must cease the processing of data subjects' data. At that point, data subjects can request the erasure of their data pursuant to art. 17, § 1, c.

Art. 21, § 6, GDPR disciplines the exercise of the right to object in the case in which data are processed for research purposes. Also in this case, data subjects have the right to object to the processing "on grounds relating to [their] particular situation"[215]. The data subject's objection to the processing can be overcome only where the controller demonstrates that the processing is necessary for the performance of a task carried out for reasons of public interest. The text of 21, § 6, poses some interpretative challenges.

First, art. 21, § 6, does not contain an express reference to the legal bases indicated in art. 21, § 1. According an interpretation in line with the rationale of the right to object, art. 21, § 6, finds application where the legal bases of processing are those indicated in art. 21, § 1, i.e., legitimate interest and processing necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller. As said, art. 21, § 6 provides that data subjects' objection to processing could be overcome by the controller only where "the processing is necessary for the performance of a task carried out for reasons of public interest". This seems to imply that, in cases in which the legal basis of the processing is legitimate interest, controllers could not be able to overcome data subjects' objection. Another difficulty is posed by the fact that art. 21, § 6 does not expressly indicate the legal effect of the successful exercise of the right to object. In line with art. 21, § 1, such effect seems to be that controllers shall no longer process the personal data. Following this interpretation, data subjects whose data are processed for scientific research purposes on the basis of a legitimate interest would be granted with a power to determine the ceasing of processing higher than the cases in which the purpose of the processing is not scientific research[216]. At the same time, the understanding of the concrete effects of art. 21, § 6, is made difficult by the interpretation of the connection between the latter and art. 17[217]. Art. 17, § 1, c, grants data subjects with the right to obtain the erasure of their data where they have successfully exercised the right to object *pursuant to art. 21, §§ 1 and 2*. That is, art. 17, § 1, c, does not contain a reference to art. 21, § 6. This interpretative obstacle can be overcome by considering the general rationale of the right to objects established by art. 21, thereby recognising that art. 17, § 1, c should be read as referring also to art. 21, § 6. However, a further interpretive difficulty is posed by the derogatory provisions established by art. 17, § 3. In particular, art. 17, § 3, d, provides that the general discipline established by art. 17, §§ 1 and 2 shall not apply to the extent that:

- the processing is necessary for scientific research purposes;
- such processing respects the requirement of art. 89, § 1;
- the right to erasure is likely to render impossible or seriously impair the achievement of the objectives of that processing.

The interpretation of such provision, *in se*, and in connection with art. 21, § 6, GDPR, is challenging. Where the derogation provided by art. 17, § 3, d, finds application, it seems that the further consequence of the successful exercise of the right to object pursuant to art. 21, § 6, could not be the erasure of the data.

More in general, the effect of the derogatory provision established by art. 17, § 3 seems to be that of making lawful a processing that infringes the GDPR or EU or national law[218]. However, in most of the cases indicated by art. 17, § 3, this effect is only apparent. The further retention of data is not made lawful by art. 17, § 3, but by the specific ground that underlies the processing indicated at the letter a, b, c, e, of art. 17, § 3. For instance, art 17, § 3, a) and e) concern a form of processing that is inherent to the exercise of a fundamental right, i.e., freedom of expression and information, and the right to establish, exercise or defend a legal claim. In the cases indicated art. 17, § 3, b) and c), the

[215] Art. 21, § 6, GDPR

[216] Such power would be even wider where the lack of indication of legal bases in the text of art. 21, § 6 was interpreted as making the right to object thereby disciplined applicable to all processing performed for a scientific research purpose, irrespective of the legal basis of processing.

[217] And art. 18, § 1, d, GDPR

[218] Recital 65, GDPR: "A data subject should have … a 'right to be forgotten' where the retention of such data infringes this Regulation or Union or Member State law to which the controller is subject. …However, the further retention of the personal data should be lawful where it is necessary, … for … scientific … research purposes … purposes …"

lawfulness of the retention of data depends on the fact that the latter is necessary to comply with a legal obligation, to perform a task carried out in the public interest or in the exercise of official authority vested in the controller, or for reasons of public interest in the area of public health. Within this group of exceptions, scientific research seems to be an outsider, since it not necessarily related to the exercise of a fundamental right nor to a legally mandated task.

In this sense, the exception to the right to erasure established by art. 17, § 3, d, seems to contrast with the requirement of lawfulness of processing. This is especially the case with respect to the provision of art. 17, § 1, b, i.e., the case in which data subjects request the erasure of their data after having withdrawn the consent to the processing. It should be noted that 17, § 1, b, concerns not only the withdrawal of consent as legal basis of the processing pursuant to art. 6, § 1, a, but also the withdrawal of consent as a ground to lift the prohibition of processing particular categories of data pursuant to art. 9, § 2, a. In such cases, the only interpretation of art. 17, § 3, d, coherent with the GDPR as a whole is that according to which the right to erasure can be derogated only where the lawfulness of the processing is ensured by the circumstance that the controller can rely on a legal basis and/or an exception to the prohibition established by art. 9 further than the consent of the data subject. In line with the considerations made in relation to further processing[219], it seems that such further legal basis and exception pursuant to art. 9, § 2, should pre-exist the exercise of the right to erasure. Where controllers could simply switch to another basis/exception on the occasion of the exercise of the right to erasure, the providing of consent by data subjects would be deprived of meaning, with the frustration of the latter's legitimate expectations. Assuming that this interpretation is correct, however, one could argue that the processing was never really based on consent, and thereby not made unlawful by its withdrawal. As a consequence, in this case the derogation provided by art. 17, § 3, d, would merely make explicit an implication that could have already be drawn from the other provisions of the GDPR.

Major difficulties are posed by the interpretation of the combination of art. 17, § 3, d and art. 17, § 1, d. According to the latter provision, data subjects have the right to obtain from controllers the erasure of their personal data - and controllers have the obligation to erase personal data without undue delay - where the personal data have been unlawfully processed. Arguably, the mere purpose of processing, even where it is scientific research, cannot by itself remedy the eventual unlawfulness of processing. At the same time, unlawful processing still subject to corrective measures by the DPA such "a temporary or definitive limitation including a ban on processing[220] and the imposition of administrative fines[221].

Additional uncertainty – and space for interpretation – results from the phrasing of art. 17, § 1, and art. 17, § 3, d. The former distinguishes i) *the right of data subjects obtain from controllers the erasure of their personal data* and ii) *the obligation of controllers to erase personal data without undue delay.* Arguably, the obligation of controllers could be understood as a duty, i.e., an obligation that applies independently from the request of data subjects. This interpretation would be coherent with the core provisions of the GDPR, which require controllers to take all measures necessary to ensure compliance with data protection law. Processing or continuing the processing, also merely in the form of storing, of personal data in the circumstances indicated by art. 17, § 1 – i.e., where such data are not anymore necessary, or where consent has been withdrawn, or where the processing is unlawful, etc. etc. - would clearly violate controllers' obligations under the GDPR. This implies that controllers should proactively proceed to the erasure of data without necessarily waiting for the request of data subjects. Otherwise, controllers would have an incentive to not inform data subjects of the processing. Even without necessarily assuming that the obligation referred to in art. 17, § 1, is a duty, it can be noticed that art. 17, § 3, d, allows the derogation of art. 17, §§ 1-2 "in so far *as the right referred to in paragraph 1* is likely to render impossible or seriously impair the achievement of the objectives of that processing". That is, art. 17, § 3, d, does not make reference to a derogation to controllers' "obligation to erase personal data without undue delay". None of the interpretive options made available by the

[219] *Supra*, A.2.2.2.

[220] Art. 58, § 2, f, GDPR. Art. 58, § 2, g, GDPR, grants DPAs with the power "to order the rectification or erasure of personal data or restriction of processing pursuant to Articles 16, 17 and 18 and the notification of such actions to recipients to whom the personal data have been disclosed pursuant to Article 17(2) and Article 19". The relation between such provision and the exception established by art. 17, § 3, GDPR, is unclear. In particular, it is unclear whether the application of art. 17, § 3, GDPR implies the restriction of the power of DPAs to order the erasure of data. In such a case, doubts similar to those illustrated above arise with respect to the combination of art. 17, § 1, d and art. 17, § 3, d.

[221] Art. 83 GDPR

text of art. 17 seems conclusive. So far, no clarification has been provided by the EDPB and the EDPS.
At this stage, we can point out that considerations similar to those made with reference to the derogation of information obligations should apply also to the provisions of art. 17 and 21. Also in these cases, the derogations to the general discipline shall be interpreted narrowly. Controllers who intend to avail themselves of the derogations bear the burden to prove the existence of the requirements demanded by the law. This applies in particular to the proof that the erasure of the data would render *impossible* or *seriously impair* the achievement of the objectives of the research. Such requirement should be read in conjunction with the requirement to respect the obligations set by art. 89, § 1, GDPR. As discussed *supra*, art. 89, § 1, requires controllers to adopt the highest degree of pseudonymisation possible. In this sense, the derogation to the right to erasure provided by art. 17, § 3, d, could be understood as a special case that follows the same rationale of art. 11. That is, the derogation to the right/obligation to erasure would be justified in the light of the measures of pseudonymisation adopted by controllers. In such case, controllers can demonstrate that they are not "in a position to identify the data subject"[222] and, therefore, they are unable to single out the data that should be erased. If this interpretation is correct, art. 17, § 3, d, would allow controllers to refuse an erasure request also when, albeit potentially capable of identifying the data subject, controllers can demonstrate that the erasure of data would negatively affect the research in a significant way.
Another aspect to be considered concerns the combined effects of the exception established by art. 17, § 3, d, and the principle of segregation illustrated in section A.1.. The exception to the provisions of art. 17, §§ 1 and 2, can only apply to processing performed for scientific research purposes. The requirements established by such provisions would therefore expand back as soon as the data are processed for purposes other than scientific research. This circumstances requires researchers that enjoy the exception to adopt a set of data management measures, especially if they intend to disseminate the data. At least, the recipients who want to use the data for purposes other than scientific research should be informed of the circumstance that the lawfulness of the processing of such data depends on the special requisites established by art. 17, § 3, GDPR. This is especially relevant in the case in which the exception established under the special regime is used to remedy an unlawful processing of data[223].

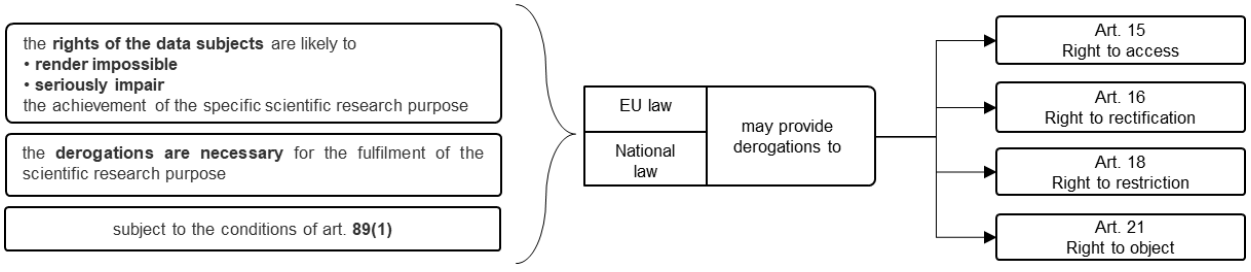
A.2.3. The third layer: EU and Member State Law

In this section, we briefly illustrate the provisions of the GDPR that empower the European and Member States' legislator to introduce derogatory provisions for the processing of personal data performed for scientific research purposes.

A.2.3.1. Safeguards and derogations ex art. 89, § 2. GDPR

As anticipated above, next to the general provision of art. 89, § 1, and the specific derogations contained in single provisions of the GDPR, art. 89, § 2, establishes a set of conditions under which the EU and Member States law can introduce further derogations that apply to processing performed for scientific research purposes. The scope of the further derogations that can be introduced by EU and national law is limited to some of the rights of data subjects, as illustrated in the figure below.

Figure A.4.: Safeguards and derogations ex art. 89, § 2. GDPR



[222] Art. 11 GDPR
[223] i.e., by combining art. 17, § 3, d, Recital 65 and art. 17, § 1, d, GDPR

It is out of the scope of the present report to give an account of the provisions adopted at the EU and Member States level pursuant to art. 89, § 2. In any case, it is worth highlighting the relevance of the conditions listed in art. 89, § 2. As the GDPR empowers EU and national legislators to introduce derogations, it also limits such law-making power. The derogations introduced by the EU and national legislators are legitimate only to extent that they satisfy the conditions set by art. 89, § 2, GDPR. This means that, in reading the derogatory provisions introduced by national or EU law, controllers should be aware of the overarching normative framework within which such derogations operate. The derogatory provisions introduced by national or EU law might be incompatible with GDPR or unconstitutional and, therefore, be disapplied in legal proceedings or be declared invalid.

A.2.3.2. The exception to the prohibition of processing special categories of data (art, 9, 2, (j))

Art. 9, § 2, j, GDPR contains another derogatory provision the application of which depends on EU or Member State law. Art. 9, § 1, GDPR prohibits the processing of particular categories of data, i.e.:

- personal data revealing
    - racial or ethnic origin
    - political opinions
    - religious or philosophical beliefs
    - trade union membership
- genetic data,
- biometric data for the purpose of uniquely identifying a natural person,
- data concerning health
- data concerning a natural person's sex life or sexual orientation.

Such prohibition can only be overcome in the cases provided by art. 9, § 2. Amongst such exceptions, art. 9, § 2, j, makes reference to the cases in which "processing is necessary for … scientific … research purposes … in accordance with Article 89, § 1, based on Union of Member State law". The reference to EU and Member States law implies that researchers interested in availing themselves of such derogation must identify a legal ground outside of the GDPR that lift the prohibition established by art. 9, § 1. Where the exception is established in national law, researchers should pay particular attention in cases in which the research is conducted across the borders of EU Member States. The exception is likely to have a scope of application limited to the processing activities performed on the territory of the State. At the same time, it cannot be taken for granted that other States will have comparable derogatory provisions.

Moreover, also in the case of the exception provided by art. 9, § 2, j, the validity of the law establishing the exception is conditioned to the respect of the conditions set by the GDPR. Pursuant to art. 9, § 2, j, the EU or national law providing the exception must:

- be proportionate to the aim pursued;
- respect the essence of the right to data protection;
- provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.

For the same reasons illustrated with respect to art. 89, § 2, researchers who want to avail themselves of the exception to the prohibition set by art. 9, § 1, are encouraged to carefully check whether the EU and national law providing the exception satisfy the conditions provided by art. 9, § 2, j, GDPR.

A.3. The limits of the special regime: which obligations *are not* derogated under the research regime

As illustrated in the previous paragraphs of the Annex and in sections 1.2.2. and 2 of the Report, the special regime established by the GDPR for processing performed for scientific research purposes does not consist only in derogatory provisions. The carrying out of processing for scientific research purposes attributes to researchers also specific obligations and additional safeguards that might not be necessary in case of other purpose of processing.

At the same time, we have seen that the derogations established by the special regime with respect to the general discipline of the GDPR concern primarily the provisions on data subjects rights. The applicability of such derogations, in turn, depends on controllers' implementation of measures aimed

at ensuring the respect of the principle of data minimisation, such as pseudonymisation measures. To the extent that such measures reduces the risks that the processing poses to data subjects, the requirements that controllers must fulfil with respect to data subjects' rights can be proportionally reduced. In any case, the derogations must find their justification in their strict necessity to the safeguard of the achievability of the research purpose. Next to such, as it were, *internal limits*, the scope of the special regime is constrained by *external limits*, i.e., by the requirements that the GDPR establishes for all forms of processing, irrespective of their – even scientific research – purpose[224]. Controllers that carry out processing that are unlawful for lack of compliance with the general provisions set by the GDPR cannot enjoy the effects of the derogations provided by the research regime. This, once again, stresses the need to avoid a siloed understanding of legal compliance and calls for an approach to the GDPR that understands data protection law both as a whole and as part of a broader legal system striving for integrity.

[224] See, section 1.2.2. and 2 of the Report

References

Becker R and others, 'Secondary Use of Personal Health Data: When Is It "Further Processing" Under the GDPR, and What Are the Implications for Data Controllers?' (2022) 1 European Journal of Health Law 1

Bender EM and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) <https://dl.acm.org/doi/10.1145/3442188.3445922> accessed 28 April 2023

Bender EM and Friedman B, 'Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science' (2018) 6 Transactions of the Association for Computational Linguistics 587

Benjamin M and others, 'Towards Standardization of Data Licenses: The Montreal Data License' (arXiv, 20 March 2019) <http://arxiv.org/abs/1903.12262> accessed 28 April 2023

Benjelloun O, Chen S, and Noy N, Google Dataset Search by the Numbers, in Pan JZ and others (Eds.), The Semantic Web – ISWC 2020. 19th International Semantic Web Conference Athens, Greece, November 2–6, 2020 Proceedings, Part II, Springer, 2020, pp 667–682. https://doi.org/10.1007/978-3-030-62466-8

Boyd KL, 'Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data' (2021) 5 Proceedings of the ACM on Human-Computer Interaction 438:1

Chen K and others, 'Deep Learning for Sensor-Based Human Activity Recognition: Overview, Challenges, and Opportunities' (2021) 54 ACM Computing Surveys 77:1

Ciliberto M and others, 'Opportunity++: A Multimodal Dataset for Video- and Wearable, Object and Ambient Sensors-Based Human Activity Recognition' (2021) 3 Frontiers in Computer Science <https://www.frontiersin.org/articles/10.3389/fcomp.2021.792065> accessed 28 April 2023

Ciliberto M, Ponce Cuspinera LA and Roggen D, 'Collecting a Dataset of Gestures for Skill Assessment in the Field: A Beach Volleyball Serves Case Study', *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (Association for Computing Machinery 2021) <https://doi.org/10.1145/3460418.3479355> accessed 28 April 2023

Comandè G and Schneider G, 'Differential Data Protection Regimes in Data-Driven Research: Why the GDPR Is More Research-Friendly Than You Think' (2022) 23 German Law Journal 559

Cooper AF and others, 'Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning', *2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) <https://doi.org/10.1145/3531146.3533150> accessed 28 April 2023

Cunningham W, 'The WyCash Portfolio Management System', *Addendum to the proceedings on Object-oriented programming systems, languages, and applications (Addendum)* (Association for Computing Machinery 1992) <https://dl.acm.org/doi/10.1145/157709.157715> accessed 28 April 2023

Demetzou K, 'GDPR and the Concept of Risk': in Eleni Kosta and others (eds), *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data: 13th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Vienna, Austria, August 20-24, 2018, Revised Selected Papers* (Springer International Publishing 2019) <https://doi.org/10.1007/978-3-030-16744-8_10> accessed 29 April 2023

Demrozi F and others, 'Human Activity Recognition Using Inertial, Physiological and Environmental Sensors: A Comprehensive Survey' (2020) 8 IEEE Access 210816

Denton E and others, 'On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet' (2021) 8 Big Data & Society 20539517211035956

Dodge J and others, 'Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus' (arXiv, 30 September 2021) <http://arxiv.org/abs/2104.08758> accessed 28 April 2023

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 11

Gebru T and others, Datasheets for datasets, Communications of the ACM, December 2021, Vol. 64 No. 12, Pages 86-92, DOI:10.1145/3458723

Geiger RS and others, 'Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2020) <https://dl.acm.org/doi/10.1145/3351095.3372862> accessed 28 April 2023

Gellert R, The Risk-Based Approach to Data Protection (Oxford University Press, 2020)

Harvey A and LaPlace J. Exposing.ai. https://exposing.ai, 2021

Hildebrandt M, 'Saved by Design? The Case of Legal Protection by Design' (2017) 11 NanoEthics 307

----------, *Smart Technologies and the End(s) of Law* (Edward Elgar Publishing 2015).

Holland S and others, 'The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards' (arXiv, 9 May 2018) <http://arxiv.org/abs/1805.03677> accessed 28 April 2023

Hutchinson B and others, 'Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2021) <https://dl.acm.org/doi/10.1145/3442188.3445918> accessed 28 April 2023

Kaneen CK and Petrakis EGM, 'Towards Evaluating GDPR Compliance in IoT Applications' (2020) 176 Procedia Computer Science 2989

Loideain NN, 'A Port in the Data-Sharing Storm: The GDPR and the Internet of Things' (2019) 4 Journal of Cyber Policy 178

Luccioni AS and others, 'A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication', *2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) <https://doi.org/10.1145/3531146.3533086> accessed 28 April 2023

Manis ML, 'The processing of personal data in the context of scientific research. The new regime under the EU-GDPR' [2017] BioLaw Journal - Rivista di BioDiritto 325

Miceli M and others, 'Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2021) <https://dl.acm.org/doi/10.1145/3442188.3445880> accessed 28 April 2023

Minh Dang L and others, 'Sensor-Based and Vision-Based Human Activity Recognition: A Comprehensive Survey' (2020) 108 Pattern Recognition 107561

Orr W and Davis JL, 'Attributions of Ethical Responsibility by Artificial Intelligence Practitioners' (2020) 23 Information, Communication & Society 719

Pareek P and Thakkar A, 'A Survey on Video-Based Human Action Recognition: Recent Updates, Datasets, Challenges, and Applications' (2021) 54 Artificial Intelligence Review 2259

Paullada A and others, 'Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research' (2021) 2 Patterns 100336

Peng K, Mathur A and Narayanan A, 'Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers' (arXiv, 21 November 2021) <http://arxiv.org/abs/2108.02922> accessed 28 April 2023

Petrozzino C, 'Who Pays for Ethical Debt in AI?' (2021) 1 AI and Ethics 205

Pushkarna M, Zaldivar A and Kjartansson O, 'Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI', *2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) <https://dl.acm.org/doi/10.1145/3531146.3533231> accessed 28 April 2023

Quelle C, 'The 'Risk Revolution' in EU Data Protection Law: We Can't Have Our Cake and Eat It, Too': in Leenes R, van Brakel R, Gutwirth S, and De Hert P (eds.), *Data Protection and Privacy: The Age of Intelligent Machines*, (Hart Publishing 2017)

Quinn P and Quinn L, 'Big Genetic Data and Its Big Data Protection Challenges' (2018) 34 Computer Law & Security Review 1000

Raji ID and others, 'AI and the Everything in the Whole Wide World Benchmark' (arXiv, 26 November 2021) <http://arxiv.org/abs/2111.15366> accessed 28 April 2023

Roggen D and others, 'Collecting Complex Activity Datasets in Highly Rich Networked Sensor Environments', *2010 Seventh International Conference on Networked Sensing Systems (INSS)* (2010)

Roggen D, Pouryazdan A and Ciliberto M, 'Poster: BlueSense - Designing an Extensible Platform for Wearable Motion Sensing, Sensor Research and IoT Applications'

Sambasivan N and others, '"Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI', *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2021) <https://doi.org/10.1145/3411764.3445518> accessed 28 April 2023

Suryanarayana G, Samarthyam G, Sharma T, Refactoring for Software Design smells. Managing Technical Debt (Elsevier, 2015)

Thomas SL, 'Migration Versus Management: The Global Distribution of Computer Vision Engineering Work', *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)* (2019)

Thompson DF, 'Moral Responsibility of Public Officials: The Problem of Many Hands' (1980) 74 American Political Science Review 905

van Dijk N, Gellert R and Rommetveit K, 'A Risk to a Right? Beyond Data Protection Risk Assessments' (2016) 32 Computer Law & Security Review 286

van Eechoud M, Study on the Open Data Directive, Data Governance and Data Act and their possible impact on research, European Commission Directorate-General for Research and Innovation Directorate A — ERA & Innovation Unit A.4 — Open Science, March 2022, doi: 10.2777/71619, at 23.https://eur-lex.europa.eu/legal-content/EN/NIM/?uri=CELEX:32019L1024. See also Implementation of the Public Sector Information Directive | Shaping Europe's digital future (europa.eu).

Wachter S, 'The GDPR and the Internet of Things: A Three-Step Transparency Model' (2018) 10 Law, Innovation and Technology 266

Widder DG and Nafus D, 'Dislocated Accountabilities in the AI Supply Chain: Modularity and Developers' Notions of Responsibility' (arXiv, 27 September 2022) <http://arxiv.org/abs/2209.09780> accessed 28 April 2023

Wilkinson MD and others, 'The FAIR Guiding Principles for Scientific Data Management and Stewardship' (2016) 3 Scientific Data 160018

Yadav SK and others, 'A Review of Multimodal Human Activity Recognition with Special Emphasis on Classification, Applications, Challenges and Future Directions' (2021) 223 Knowledge-Based Systems 106970

EU normative sources

Charter of Fundamental Rights of the European Union, http://data.europa.eu/eli/treaty/char_2012/oj

Consolidated version of the Treaty on the Functioning of the European Union, http://data.europa.eu/eli/treaty/tfeu_2016/oj

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) (Text with EEA relevance), http://data.europa.eu/eli/reg/2016/679/2016-05-04

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, http://data.europa.eu/eli/reg/2018/1807/oj

Regulation (EU) 2021/695 of the European Parliament and of the Council of 28 April 2021 establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination, and repealing Regulations (EU) No 1290/2013 and (EU) No 1291/2013, http://data.europa.eu/eli/reg/2021/695/oj

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), http://data.europa.eu/eli/dir/2019/1024/oj

Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information, http://data.europa.eu/eli/dir/2013/37/oj

Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, http://data.europa.eu/eli/dir/2003/98/oj

Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM/2022/496 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496

Proposal for a Directive of the European Parliament and of the Council on liability for defective products, COM/2022/495 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2022:495:FIN

Consolidated text: Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products, http://data.europa.eu/eli/dir/1985/374/1999-06-04

ECJ, Judgment of the Court (Grand Chamber), Case C-439/19, 22 June 2021, ECLI:EU:C:2021:504, §§ 127-128

Opinion Of Advocate General Szpunar, Case C-439/19, 17 December 2020, ECLI:EU:C:2020:1054

EU Institutions Policy documents

Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A European strategy for data, COM(2020) 66 final, Brussels, 19.2.2020, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52020DC0066

Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information, C/2018/2375, http://data.europa.eu/eli/reco/2018/790/oj

Council conclusions on open, data-intensive and networked research as a driver for faster and wider innovation, 9360/15, Brussels, 29 May 2015, https://data.consilium.europa.eu/doc/document/ST-9360-2015-INIT/en/pdf

Council conclusions on the transition towards an open science system, 9526/16, Brussels, 27 May 2016, https://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf

Communication from the Commission to the European Parliament and the Council, Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union, COM/2019/250 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN

European Commission, Open Science, https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en

European Data Protection Board, Article 29 Working Party, European Data Protection Supervisor

European Data Protection Board, Guidelines on Data Protection Impact Assessment (DPIA), Adopted on 4 April 2017, As last Revised and Adopted on 4 October 2017, 17/EN, WP248rev.01, https://ec.europa.eu/newsroom/just/document.cfm?doc_id=47711

European Data Protection Board, Guidelines on transparency under Regulation 2016/679, Adopted on 29 November 2017, As last Revised and Adopted on 11 April 2018, 17/EN, WP260rev.01, https://ec.europa.eu/newsroom/article29/redirection/document/51025

European Data Protection Board, Guidelines 05/2020 on consent under Regulation 2016/679, Version 1.1, Adopted on 4 May 2020, https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf

European Data Protection Board, Guidelines 07/2020 on the concepts of controller and processor in the GDPR, Version 1.0, Adopted on 02 September 2020, https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202007_controllerprocessor_en.pdf

Article 29 Data Protection Working Party, Opinion 8/2014 on the on Recent Developments on the Internet of Things, Adopted on 16 September, 201414/EN, WP 223, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp223_en.pdf

Article 29 Working Party, Statement on the role of a risk-based approach in data protection legal frameworks, Adopted on 30 May 2014, 14/EN, WP 218, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf?wb48617274=72C54532

Article 29 Working Party, Opinion 4/2007 on the concept of personal data, Adopted on 20th June 2007, 01248/07/EN, WP 136, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf

Article 29 Working Party, Opinion 05/2014 on Anonymisation Techniques, Adopted on 10 April 2014, 0829/14/EN, WP216, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Article 29 Working Party, Opinion 03/2013 on purpose limitation, Adopted on 2 April 2013. 13/EN, WP 203, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

Article 29 Working Party, Opinion 06/2013 on open data and public sector information ('PSI') reuse, Adopted on 5 June 2013, 1021/00/EN, WP207, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp207_en.pdf

European Data Protection Supervisor, Opinion 5/2018 on the proposal for a recast of the Public Sector Information (PSI) re-use Directive, 10 July 2018, https://edps.europa.eu/sites/edp/files/publication/18-07-11_psi_directive_opinion_en.pdf

European Data Protection Supervisor, A Preliminary Opinion on data protection and scientific research, 6 January 2020, https://edps.europa.eu/sites/default/files/publication/20-01-06_opinion_research_en.pdf

European Data Protection Supervisor, Study on the appropriate safeguards under Article 89(1) GDPR for the processing of personal data for scientific research, Final Report EDPS/2019/02-08, August 2021, https://edpb.europa.eu/system/files/2022-01/legalstudy_on_the_appropriate_safeguards_89.1.pdf

European Data Protection Board-European Data Protection Supervisor, Joint Opinion 03/2021 on the Proposal for a regulation of the European Parliament and of the Council on European data governance (Data Governance Act) Version 1.1, 10 March 2021, https://edpb.europa.eu/system/files/2021-03/edpb-edps_joint_opinion_dga_en.pdf

Further sources

Speech by Giovanni Buttarelli (12 April 2018), op. cit., p. 2, “only scientific research performed within an established ethical framework”

Mauritius Declaration on the Internet of Things, 36th International Conference of Data Protection and Privacy Commissioners, Balaclava, 14 October 2014, Mauritius Declaration (europa.eu);

EU Grants: Data Management Template (HE):V1.0 – 05.05.2021, https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/reference-documents;programCode=HORIZON

Constitutional provision on the freedom of research in Member States, https://fra.europa.eu/en/eu-charter/article/13-freedom-arts-and-sciences#national-constitutional-law

Project and project partners information

https://lsts.research.vub.be/

https://www.humane-ai.eu/

https://www.humane-ai.eu/project/collection-of-datasets-tailored-for-humane-ai-multimodal-perception-and-modelling/ ; D1.1: First Year Microproject Results from Workpackage 1, 2 and 3, pp. 26-27, https://www.humane-ai.eu/wp-content/uploads/2021/10/Deliverable-1.1.pdf

https://www.humane-ai.eu/workpackages/

D1.1: First Year Microproject Results from Workpackage 1, 2 and 3, pp. 26-27, https://www.humane-ai.eu/wp-content/uploads/2021/10/Deliverable-1.1.pdf

WP 6: Applied research with industrial and societal use cases, https://www.humane-ai.eu/workpackages/