Katja de Vries

A Researcher's Guide for Using Personal Data and Non-Personal Data Surrogates: Synthetic Data and Data of Deceased People

1 Introduction

In 2018, I was working as a postdoc at the IT University (ITU) in Copenhagen. Besides teaching and research, I also sometimes acted as a mediator between the Data Protection Officer[1] (DPO) and researchers, trying to match the legal requirements following from data protection law with the practical concerns that researchers face when handling data in the, often, messy realities of doing research. In early 2020, in a Preliminary Opinion on data protection and scientific research,[2] the European Data Protection Supervisor (EDPS) wrote:

> Data protection rules aim to ensure safety and transparency while minimising interference with ethical research that aim at generalisable knowledge and societal good. The GDPR serves in part to ensure accountability for such practices. There is no evidence that the GDPR itself hampers genuine scientific research.

[1] I would like to thank Rasmus Balle and Lilian Baunbæk (that is, the Data Protection Officer and Chief Information Security Officer at ITU during that period of time) for the pleasant and insightful collaboration.

[2] European Data Protection Supervisor (EDPS), *Preliminary Opinion on data protection and scientific research*, 6 January 2020.

Does the GDPR indeed not hamper research? Talking to researchers about the GDPR has given me first-hand experience of what it must feel like to be a dentist handling anxious patients: upon entering a room, I could feel the waves of GDPR-anxiety flowing towards me. While it is true that the General Data Protection Regulation 2016/679[3] (GDPR) is much more research-friendly than many researchers might think, it does require that researchers do some substantive and pre-emptive thinking about the data they plan to process in their research. They have to ask themselves fundamental questions such as: Do I really need this data? Can I do my research in a less data-intensive way? Can I use pseudonymised or anonymised data instead? How long do I need to keep this data? What is the exact purpose that the data fulfil in my research? How can I notify data subjects about the data processing? Are the technological and organisational measures that I have taken to keep the data secure appropriate for the state-of-the-art; the potential risks to the rights and freedoms of the people whose data are processed; and the nature, scope, context and purposes of processing? If such questions are taken seriously, they require some real thinking and balancing of interests, and are not exhausted by simply ticking a box. No wonder that researchers, if they have the choice, prefer using data that fall outside the scope of the GDPR and that allow them to skip the GDPR-based soul searching about their research. This contribution reaches out a hand to researchers, especially those processing data for the creation of AI-models; moreover, it takes them on a quick tour through their options for finding data to fuel their research. In section 2, I revisit the question of whether data protection is stifling AI-research. Then in section 3, I look at initiatives that the EU legislator has recently proposed to make the lives of researchers easier, while staying within the boundaries of the GDPR, notably through the concepts of 'data intermediation services' and 'data altruism' in the proposed Data Governance Act. Thereafter, in section 4 and 5, I look at two possible surrogates for personal data, which allow researchers to escape from the scope of the GDPR: data of deceased people and synthetic data. Finally, in section 6, I present some concise conclusions and pointers for the researcher suffering from GDPR-anxiety.

[3] 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 the repealing Directive 95/46/EC, OJ L 119, 4.5.2016, 1–88.

2 Is data protection stifling AI research and innovation?

All over the world, countries are fighting to stay ahead in the Artificial Intelligence (AI) race[4] and to create a climate that stimulates AI, in terms of research and innovation as well as its adoption and commercialisation. In order to create and use AI, it is indispensable to have data, preferably of good quality and in abundant amounts. For example, to create an AI or machine learning (ML) model that can separate different types of tumours on brain scans, healthy lungs from those affected by Covid or that can recognise the author of an anonymous text, one needs to have enough training data that a model can learn from. Thus, access to data is the fuel for keeping up in the AI race.

Information technology allows data to move around quickly and seemingly effortlessly. Data, however, never move in a legal vacuum. Legal regulation mandates if data are allowed to move freely or not, and under what conditions. Legal fields that decide if data movement is permitted, obligatory, conditional or prohibited include data protection, intellectual property, the right of access to public documents based on the principle of open government, the right to re-use of publicly funded information as open data, and research ethics regulations.

Research and innovation are areas that are prioritised and fostered within the EU. While access to data can be limited by rights of others, such as intellectual property or data protection rights, the EU legislator often reserves a special regime for activities that fall in the field of research and/or innovation. To illustrate this, I will name four (proposed) legal EU instruments that all give a privileged status to data used for research purposes: the Text- and Data mining exception in the Copyright Directive, platform access in the proposed Digital Services Act, the exclusion of research from the scope of the proposed AI Act, and the research exception in the GDPR.

The first example of a research exception relates to data that are protected as works by copyright (for example, tweets, pictures or drawings that contain at least a minimal trace of authorship) or as databases by the

[4] Daniel Castro & Michael McLaughlin, Who Is Winning the AI Race: China, the EU, or the United States? – 2021 Update. Information Technology & Innovation Foundation (ITIF), at: https://itif.org/publications/2021/01/25/who-winning-ai-race-china-eu-or-united-states-2021-update (published online 25 January 2021).

sui generis database-right in Database Directive 96/9.[5] Normally, such works or databases cannot be used without permission from the holder of the intellectual property right. However, Article 3(1) of the Copyright Directive 2019/790[6] contains an exception for Text- and Data mining (TDM) 'for the purposes of scientific research', where scientific research is understood as non-commercial research.[7] This means that non-commercial researchers can train AI models on protected works and databases without needing permission from the rightsholder.

The second example is the right to platform data access which can be found in Article 31(2) and (4) of the proposed Digital Services Act,[8] and gives 'vetted researchers'[9] access to data 'for the sole purpose of conducting research that contributes to the identification and understanding of systemic risks'. While this would be helpful to certain researchers, for example, those creating AI models that capture the spread of disinformation of platforms, such as Facebook or Twitter, several commentators have criticised the narrow scope of this exception and proposed that it should be broadened, both in terms of types of researchers and research.[10]

The third example is the exclusion of research from the scope of the proposed AI Act.[11] The AI Act aims to regulate AI systems that impact on

[5] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, 20–28.

[6] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and the amending Directives 96/9/EC and 2001/29/EC, OJ L 130, 17.5.2019, 92–125.

[7] Recital 12 of the Copyright Directive 2019/790. This narrow interpretation of "scientific research" is not uncontroversial. See e.g. Rossana Ducato and Alain Strowel, Ensuring text and data mining: Remaining issues with the EU copyright exceptions and possible ways out, 43 European Intellectual Property Review, 5, 2021, 322–337.

[8] Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and the amending Directive 2000/31/EC, COM/2020/825 final, 15 December 2020.

[9] According to Article 31(4) DSA, "vetted" means that researchers are 'affiliated with academic institutions, be independent from commercial interests, have proven records of expertise in the fields related to the risks investigated or related research methodologies, and shall commit and be in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request'.

[10] Paddy Leersen, Platform research access in Article 31 of the Digital Services Act. Sword without a shield? Verfassungsblog: On matters constitutional, at: https://verfassungsblog.de/power-dsa-dma-14/ (published online 7 September 2021).

[11] European Commission, Proposal for a Regulation of the European Parliament and

society and citizens, potentially in a negative way, which means that the AI Act only concerns AI systems *in practice*, that is, those that are placed on the market, into service or used (Article 1(a) and (b)). Specifically, the adjustments[12] proposed by the Council in Article 2(6) and (7) underline the special status of AI research and development:

> Article 2(6). This Regulation shall not apply to AI systems, including their output, specifically developed and put into service for the sole purpose of scientific research and development.
>
> Article 2(7). This Regulation shall not affect any research and development activity regarding AI systems in so far as such activity does not lead to or entail placing an AI.

The new Recital 12a fleshes this point out even further by saying that the AI Act should not apply to AI systems, which are used for 'the sole purpose of research and development' in order to ensure that the Act 'does not otherwise affect scientific research and development activity on AI systems' but 'that any other AI system that may be used for the conduct of any research and development activity should remain subject to the provisions'. Thus, the Council's adjustments show that the European legislator feels the need to stress that the regulation of certain risky AI practices should not hinder AI systems with the sole purpose of research and development.

In all the aforementioned research exceptions, the research that is protected is somehow limited. The TDM-exception on copyright and database rights is limited to non-commercial science, the platform access exception only can be used by 'vetted' academics that study 'systemic risks' related to the platform data, and the AI Act only excludes AI systems that are completely disconnected from practice, whose sole purpose is research and development. In comparison, the 'scientific research' ex-

of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, COM(2021) 206 final, Brussels, 21 April 2021.

[12] Council of the European Union, 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 – Presidency compromise text, 2021/0106(COD), Brussels 29 November 2021.

ception in Art. 89 of the GDPR is much wider in scope, 'including for example technological development and demonstration, fundamental research, applied research and privately funded research' (Recital 159). The question arises: how broad is the scope of the scientific research exception exactly? This question has become even more important since the GDPR has come into force. In contrast to its predecessor, Data Protection Directive 95/46[13], the GDPR includes a scientific research exception for the processing of *sensitive* personal data (Article 9(2)j GDPR), such as data relating to health or race. Do practices that mix commercial exploitation with research also qualify as 'scientific research', in the meaning of the GDPR? Scandals like the NHS/DeepMind deal from 2016[14] raise the question as to what kind of research should benefit from the GDPR exception. Should companies that can mix treatment, research and commercial development of health-related AI, such as *Google Health*, fall under the privileged scientific research regime of the GDPR? In a recent preliminary Opinion, the European Data Protection Supervisor (EDPS) does not exclude commercial research as such but argues that the scope of the exception in the GDPR should be limited to research that is 'set up in accordance with relevant sector-related methodological and ethical standards', which includes 'the notion of informed consent, accountability and oversight' and that 'is carried out with the aim of growing society's collective knowledge and wellbeing, as opposed to serving primarily one or several private interests'.[15] Even if one follows this narrower reading of the scope of the scientific research, the scope is still very broad in comparison to the other research exceptions discussed earlier in this section. Certain types of research performed by a commercial company like *Google Health* might very well fall within the EDPS definition of 'scientific research'. However, research that is only commercial and does not follow relevant ethical or medical standards, will fall outside.

[13] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23 November 1995, 31–50.

[14] Julia Powels, Why are we giving away our most sensitive health data to Google? The Guardian, at: https://www.theguardian.com/commentisfree/2017/jul/05/sensitive-health-information-deepmind-google (published 5 July 2017); Julia Powles & Hal Hodson, Google DeepMind and healthcare in an age of algorithms. 7 Health and Technology, Issue 4, 2017, 351–367.

[15] EDPS, *A Preliminary Opinion* (n. 2), 11–12.

The broad understanding of 'scientific research' in the GDPR is all the more important because the scope of the GDPR itself is very large: personal data, that is 'any information relating to an identified or identifiable natural person' (Article 4(1) GDPR[16]), is a category of data that is extremely broad, as underlined, for example, by Purtova.[17] Even data that, at first sight, do not seem to be personal can still qualify as personal data if there is a reasonably likely potential that the data could be retraced to a living individual: I might not *directly* recognise a person from an IP address, a movement pattern or a brain scan, but with some additional research and by combining data from other data, I would be able to connect the dots. This implies that an enormous amount of research falls within the scope of the General Data Protection Regulation (GDPR) 2016/679.[18] However, in order to establish if the GDPR creates a hindrance for AI research, it is not enough to look at the scope of the exception. It is the content of the exception in Article 89 GDPR that is most crucial. The research exception entails that researchers fall under a lighter regime of data protection and have to comply with fewer requirements. Especially, the informational rights of data subjects (right to have data rectified, to access the data, right to object to the processing or to restrict it, right to erasure, etc.) are much more limited or sometimes even non-existent when it can be shown that the exercise of such rights interferes with the research. Yet, as mentioned in the introduction, the data protection requirements that have to be complied with within this privileged regime still compel substantive, pre-emptive thinking about matters like data minimisation and purpose limitation and this often instils a sense of GDPR anxiety in researchers.[19] Does this mean, notwithstanding that the research exception in the GDPR is much more generous in scope than any of the other exceptions mentioned above, that the EU has created a hurdle for itself, and that innovation might be stifled by data protection requirements? While some argue that this is the

[16] GDPR 2016/679 (n. 3).

[17] Nadezhda Purtova, The law of everything. Broad concept of personal data and future of EU data protection law, 10 Law, Innovation and Technology, 2018, 40–81.

[18] GDPR 2016/679 (n. 3).

[19] Katharina Ó Cathaoir, Hrefna Dögg Gunnarsdóttir & Mette Hartlev, The journey of research data: Accessing Nordic health data for the purposes of developing an algorithm, *Medical Law International*, 2021, 1–23.

case,[20] others consider GDPR anxiety to be a result of misunderstandings and hence a transitional matter – the requirements are not unreasonably burdensome, and it simply takes a while to get used to GDPR compliance. Here, an analogy with environmental law[21] could be made, as the concern that the GDPR will hinder innovation resembles the ones raised during the 1970s and 80s about how environmental laws could turn out to be extremely detrimental for businesses and international competitiveness. These concerns about environmental regulation have been refuted,[22] and the benefits of regulating an industrial wild west outweigh the costs: if left unregulated, the behaviour of industry is likely to result in an environmental tragedy of the commons. The same arguments are raised to defend data protection and the regulation of AI.[23] While the analogy between environmental and data protection works well in many respects, there are at least two important differences between the latter and the former. Firstly, the number of actors that is affected by data protection is much larger. Data protection does not only affect big companies like *Amazon*, *Google* or *Meta*; it also affects any individuals or clubs who publish information about other people on their private website, the local

[20] Tal Zarsky, Incompatible: the GDPR in the age of big data, 47 Seton Hall L. Rev., 2016, 995–1020.

[21] This is an analogy that has been made quite frequently in academic literature, although the main point of comparison tends to be that the regulation of data protection and environmental protection use similar regulatory tools (risk impact assessments, transparency requirements, etc.) and that a lack of regulation leads to a tragedy of the commons in both fields. See e.g.: Magdalena Słok-Wódkowska & Joanna Mazur, Regulating the digital environment: what can data protection law learn from environmental law?, 19 Review of International, European and Comparative Law, 2021, 13–43; Mary Julia Emanuel, Evaluation of US and EU Data Protection Policies Based on Principles Drawn from US Environmental Law in: D. Svantesson & D. Kloza (eds), Trans-Atlantic Data Privacy Relations as a Challenge for Democracy, Cambridge 2017, 407–427; A. Michael Froomkin, Regulating Mass Surveillance as Privacy Pollution: Learning from Environmental Impact Statements, University of Illinois Law Review, Issue 5, 2015, 1713–1790; Dennis D. Hirsch, Protecting the Inner Environment: What Privacy Regulation Can Learn from Environmental Law, 41 Georgia Law Review, Issue 1, 2016, 1–63.

[22] Frank S Arnold, Anne S Forrest & Stephen R Dujack, *Environmental Protection: Is it Bad for the Economy?*: Environmental Law Institute 1998.

[23] See, for example, the extensive report produced by the Dutch Scientific Council for Government Policy: Corien Prins, Haroon Sheikh, Erik Schrijvers, Eline De Jong, Monique Steijns & Mark Bovens. *Mission AI. The New System Technology. Summary report*, Wetenschappelijke Raad voor het Regeringsbeleid (Scientific Council for Government Policy), The Hague, Netherlands, 2021.

supermarket that has a customer loyalty programme, the municipality dealing with citizens, researchers doing research involving personal data, etc. The list is, in principle, endless, and it is currently difficult to find anyone inside the EU who has not heard or been affected by the GDPR. Even if I only focus on personal data processing in *AI research*, which is only a tiny domain within the much larger material scope of the GDPR, the impact of the GDPR is enormous in comparison to environmental regulation of industry. This brings me to a second difference, which I have already discussed above: namely that even under the lighter regime of the research exception (Article 89 GDPR), the GDPR still requires individual researchers to do some fundamental and substantive thinking about the data needed for their research. Compared to a data protection impact assessment (looking at the risks to individual *rights*),[24] an environmental impact assessment (looking at the risks to the *environment*) has a more tangible and quantifiable object Loss in biodiversity or increase in CO_2 can be quantified. The numbers might be up for debate but at least some quantification is possible. Quantifying the risk to a right is more difficult: how to put a number to how much privacy is lost, how much more cautious citizens become due to an increased chilling effect, or how much of the rule of law evaporates? Researchers processing personal data will often have to do some balancing of interests in light of a deep understanding of their research and its impact. A DPO can *assist* a researcher in asking the right GDPR questions, such as 'Is the public interest pursued important enough to legitimise the negative effects for affected individuals?' However, the answer to these questions can only be found by *combining* detailed knowledge about the research set-up and purposes with an understanding of data protection law. Of course, one should not exaggerate the burden of data protection compliance. Thinking seriously about research data (how you use them, how long you need to keep them, why you need them, how to keep them secure, how you notify affected data subjects, etc.) should, in principle, not be an insurmountable burden. Yet, in the busy day-to-day life of many researchers, the discovery that GDPR-compliance requires more than the thoughtless ticking of a few boxes, combined with the fear of hefty administrative fines, in the case of non-compliance (Article 83 GDPR), can cause the

[24] Raphael Gellert, *The Risk-based Approach to Data Protection*, Oxford University Press (2020).

aforementioned GDPR anxiety. Moreover, some researchers[25] might argue that notably the principle of *purpose specification* in Article 5(1)(b) GDPR, which requires personal data to be collected for a 'specific, explicit and legitimate' purpose, and *data minimisation* in Article 5(1)(c), which requires that data be 'limited to what is necessary in relation to the purposes for which they are processed', are at odds with the flexible attitude underlying much big data research that boils down to 'Let's gather as much data as I can, and then just try out some things – I'll find out by trial and error what generates interesting results'.

3 How the EU legislator wants to make the life of researchers easier: Data intermediation services and data altruism

Is GDPR anxiety just another name for a sloppy research attitude, entailing the lack of proper hypotheses and research plans, and a too limited understanding of the opportunities offered to researchers in the GDPR? It might be – in some cases, at least – but the EU legislator clearly feels the need to help researchers by making data processing within the boundaries of the GDPR easier. One of the proposals that could help to realise these ambitions is the proposed Data Governance Act[26] (DGA), presented by the Commission in November 2020. On 31 November 2021, the European Parliament and Council reached an agreement and presented a provisional final version.[27] In Recital 5 of this latest version of the proposed DGA, it states in relevant part:

[25] Zarsky (n. 20).

[26] European Commission, Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act), COM(2020) 767 final, Brussels, 25 November 2020. It should be noted that the DGA not only tries to facilitate the sharing of personal data, but also data which are protected by intellectual property rights. In this contribution, I only focus on the sharing of personal data in the DGA, but the DGA mechanisms for both categories of data are more or less the same.

[27] Council of the European Union, Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act) – Analysis of the final compromise text in view to agreement, 2020/0340(COD), Brussels, 10 December 2021.

> …certain categories of data (commercially confidential data, data subject to statistical confidentiality, data protected by intellectual property rights of third parties, including trade secrets and personal data) in public databases is often not made available, despite this being possible in accordance with the applicable Union law, in particular Regulation (EU) 2016/679 and Directives 2002/58/EC and (EU) 2016/680, not even for research or innovative activities in the public interest. Due to the sensitivity of those data, certain technical and legal procedural requirements must be met before they are made available, not least in order to ensure the respect of rights others have over such data, or limit negative impact on fundamental rights, the principle of non-discrimination and data protection. Such requirements are usually time- and knowledge-intensive to fulfil. This has led to the underutilisation of such data (…) In order to facilitate the use of data for European research and innovation by private and public entities, clear conditions for access to and use of such data are needed across the Union.

The DGA basically introduces three trajectories to incentivise sharing of data that is protected by data protection or intellectual property rights. Firstly, it gives guidance on the re-use of protected data owned by public sector bodies (Chapter II, DGA). Secondly, it creates a "data altruism" framework (Chapter IV, DGA), facilitating the sharing of protected data for the common good, including research. And finally, it introduces a framework for so-called "data intermediation services" (Chapter III, DGA), that is, professional data sharing services. One central idea in the DGA is to create sector-specific "data spaces", which could be described as data silos or commons, managed by the aforementioned data intermediaries. When data are kept in such a data space, supervised and managed by a professional intermediary, this would hopefully lead to improved data quality, reliability, availability and security of data, which would automatically also entail a higher level of GDPR compliance and public trust, as well as a more streamlined and institutionalised process for requesting permission to use the data. Some of these data spaces, such as the European Health Data Space, will require additional regulation[28] because of the specific sensitive nature of certain types of data and particular sectorial demands. The DGA will, however, provide the common

[28] Towards European Health Data Space (TEHDAS), *Milestone 5.8 Potential health data governance mechanisms for European Health Data Space*, 1 September 2021, project report co-funded by the European Union's 3rd Health Programme (2014–2020) under Grant Agreement no 101035467.

framework. In Article 2(c) of the DGA, data intermediation service is defined as 'a service, which aims to establish commercial relationships for the purpose of data sharing between an undetermined number of data subjects and data holders, on the one hand, and data users on the other hand, through technical, legal or other means, including for the exercise of data subjects' rights in relation to personal data'.[29] Data intermediaries will form a new commercial business model which excludes existing non-profit collaborative knowledge platforms (such as *WikiMedia*), as well as commercial businesses that only provide a technical means for sharing without establishing a legal and commercial relationship between potential sharers and users (such as cloud services like *OneDrive* and *Dropbox*):

> The provision of cloud storage, analytics or of data sharing software, the provision of web browsers or browser plug-ins, or an email service should not be considered data intermediation services in the sense of this Regulation, as long as such services only provide technical tools for data subjects or data holders to share data with others, but are neither used for aiming to establish a commercial relationship between data holders and data users, nor allow the provider to acquire information on the establishment of commercial relationships for the purpose of data sharing, through the provision of such services. Examples of data intermediation services would include, inter alia, data marketplaces on which companies could make available data to others, orchestrators of data sharing ecosystems that are open to all interested parties, for instance in the context of common European data spaces, as well as data pools established jointly by several legal or natural persons with the intention to license the use of such pool to all interested parties in a manner that all participants contributing to the pool would receive a reward for their contribution to the pool. This would exclude value-added data services, that obtain data from data holders, aggregate, enrich or transform the data for the purpose of adding substantial value to it and license the use of the resulting data to data users, without establishing a commercial relationship between data holders and data users.[30]

The idea is that data intermediaries would be registered, supervised by a new supervisory body called the 'European Data Innovation Board' and easily recognisable through a common logo that identifies them as a provider of 'data intermediation services recognised in the Union'. As such, these intermediaries, who *only* act as intermediaries and not use the

[29] DGA-Council (n. 28), Article 2c.

[30] DGA-Council (n. 28), Recital 22a.

data themselves for other purposes (Article 2(c) of the DGA), would offer natural and legal persons an alternative to simply parking their data at some integrated tech platform. Storing data at an intermediation service would offer a way for data subjects and data holders to stay in control of the data connected to them, through data protection or intellectual property rights, while at the same time allowing for data sharing for certain purposes.

One important adjustment in the latest version of the DGA is a clarification in Recital 3a, namely that the GDPR has an incontestable primacy over the DGA:

> This Regulation should in particular not be read as creating a new legal basis for the processing of personal data for any of the regulated activities, or as modifying information requirements under Regulation (EU) 2016/679.[31]

Thus, if the intermediation services are to make data sharing easier, this would not be because the regulatory data protection regime is altered. The introduction of intermediation services in the DGA aims to be like a *Tinder* of sharing data. Without altering the data protection rules as such, the hope is that the introduction of data intermediaries, who match potential data sharers with potential data (re-)users, could be as much of a game changer as *Tinder*-like services were for dating.

In order to further incentivise natural and legal persons to share data with data, the EU legislator has also introduced the concept of 'data altruism' in Article 19 DGA. In contrast to data intermediaries, data altruism organisations and the natural or legal persons sharing their data for altruistic purposes do this on a *non-profit* basis. Data altruism organisations should, like data intermediaries, be recognisable by a common logo. Article 2(10) of the DGA states that 'data altruism' amounts to 'voluntary sharing of data based on consent by data subjects to process personal data pertaining to them, or permissions of other data holders to allow the use of their non-personal data without seeking or receiving a reward that goes beyond a compensation related to the costs they incur making their data available, for purposes of general interest'. As pointed out by González Fuster,[32] the choice of the word 'data altruism', instead of, for example, the more neutral term 'data donation', gives a strong normative value to

[31] DGA-Council (n. 28), Recital 3a.

[32] Gloria González Fuster, Carta Academica. L'altruisme des données peut-il sauver le monde? Le Soir, 24 April 2021.

the concept. 'Data altruism' has a morally positive ring to it, whereas its opposite, 'data egoism', sounds less appealing. Moreover, González Fuster continues, the problem with the word 'data donation' is also that data protection is understood in the EU as an inalienable fundamental right that should not be understood in terms of property rights. One cannot sell or give away one's right to data protection, in the same way as one cannot do that with other inalienable rights, such as one's right to human dignity. The word 'data altruism' makes it easier to defend that data are not donated, in the meaning of a transfer of property, but that the data subject consents to its use in compliance with the GDPR. Not unlike the data spaces managed by commercial intermediaries, data altruism organisations fulfil the role of a dating market between potential data sharers and users, but what brings them together is a non-commercial shared commitment to a particular purpose of general interest. Article 22 of the proposed DGA offers the possibility to the Commission to adopt implementing acts for the development of a uniform European data altruism consent form, using a modular approach, allowing customisation for specific sectors and for different purposes. This consent can, in line with the GDPR, be revoked at any point. It could, however, be questioned if the data altruism in the proposed DGA is as fully GDPR compatible, as it claims to be.[33] In its position paper[34] on the DGA, the European Consumer Organisation (BEUC) warns that the term 'purposes of general interest' (Article 2(10) of the DGA) is too vague. The term can easily be stretched in unforeseeable ways:

> Consumers must also be legally protected against misleading practices which are presented as public purpose research when in reality there is commercial intent in the exploitation of the data as a result of the commercialisation of the research outputs.[35]

[33] Paul Keller and Francesco Vogelezang, The Data Governance Act – between undermining the GDPR and building a Data Commons, *EDRI*, at: https://edri.org/our-work/the-data-governance-act-between-undermining-the-gdpr-and-building-a-data-commons/ (published online 14 July 2021); Paul Keller and Francesco Vogelezang, The Data Governance Act: five opportunities for the data commons, *Open Future*, at: https://openfuture.eu/publication/the-data-governance-act-five-opportunities-for-the-data-commons/ (published online 23 June 2021).

[34] The European Consumer Organisation (BEUC), *Data Governance Act. BEUC position paper*, 2021.

[35] BEUC (n. 35), p. 3.

In the initial version of the DGA, 'general interest' was left undefined and only exemplified by two examples in Article 2(10): '…such as scientific research purposes or improving public services'. BEUC criticised this lack of a definition of 'general interest' in the DGA and wrote that:

> there are no clear legal benchmarks to check against the presence of such a 'general interest' ('altruism washing') and, in some cases, the interpretation of what constitutes a 'general interest' might differ at national level.[36]

In the latest version of the DGA, 'general interest' is illustrated with more examples in Article 2(10); nonetheless, at a fundamental level, BEUC's criticisms regarding vagueness still hold:

> …for purposes of general interest, defined in accordance with national law where applicable, such as healthcare, combating climate change, improving mobility, facilitating the establishment of official statistics, improving public services, public policy making or scientific research purposes in the general interest.

It would be up to altruism organisations or the data receiving public entity to ensure that the altruistically shared data are shared for a purpose or set of purposes that can be qualified as 'general interest' and that are sufficiently specific to be in accordance with the purpose specification principle in Article 5(1)(b) of the GDPR. The importance of compliance with the purpose specification principle is clarified by a new addition in Article 19(1)a. In the initial version of this Article, it only said that data altruism organisations should inform data holders 'about the purposes of general interest for which it permits the processing of their data by a data user', whereas the latest version also specifies that information about 'the specified, explicit and legitimate purpose' for which it permits the processing of personal data should be provided. However, the fact that a data space, managed by a data intermediation service or data altruism organisation, is supposed to be a one-stop shop, where a multitude of actors can request access to data, seems to create an incentive to not make the purposes too specific, and makes it attractive to stretch out the specificity of purposes to the maximal vagueness still permitted by the GDPR. Moreover, in the case of data altruism, one could basically imagine two different scenarios with regard to purpose specification. The first one is

[36] BEUC (n. 35), p. 8.

where potential data sharers have a very specific purpose in mind, such as in the case where a patient suffering from a rare disease wants to stimulate research only in this very particular field. The second scenario is a potential data sharer who simply wants to get a quick fix of 'do-gooder' feeling and is nudged to share data for a default set of rather broadly formulated general interest purposes. The question is if the latter would cause friction with the requirements of purpose specification and freely given consent in the GDPR. Moreover, data altruism might give the false impression to data subjects that re-use of data for a new purpose always requires their renewed consent (Article 6(4) of the GDPR), whereas, in fact, this is not the case for 'archiving purposes in the public interest, scientific or historical research purposes or statistical purposes' (Article 5(1)b)) that are presumed to be compatible, as well as for re-use that is in accordance with the law, necessary in a democratic society and in pursuance of a legitimate aim (Article 6(4) of the GDPR).[37] A potential do-gooder might be surprised to find out that a planned act of altruism is void because the data already have been shared, and that the GDPR, in fact, does not always require consent for data re-use.

Does the DGA help a researcher who is experiencing GDPR anxiety? The DGA might help, in terms of data *accessibility*, in the same way a dating service like *Tinder* increases the amount of potential individuals to date. However, given that the GDPR has primacy over the DGA, the burden of GDPR compliance will not disappear. The procedures followed by data intermediation services and data altruism organisations might be more standardised, but the substantive thinking about data protection requirements cannot be removed. Nor are GDPR requirements like data minimisation and purpose specification, which might frustrate a researcher who would have the freedom to freely change between research purposes: if a data set containing gait and facial expressions of individuals in public transport does not lead to a good AI-model to identify Covid-19 infections, why not try to see if the data can be used to spot people who don't have a bus ticket or illegal migrants? Even though the GDPR, in principle, allows for jumping from one research purpose to another (compatible purpose, Article 5(1)b of the GDPR), the researcher would have to do the exercise in substantive GDPR-thinking before each shift in research purpose. In order to truly stop worrying about the

[37] Merel Koning, *The purpose and limitations of purpose limitation*, Doctoral dissertation Radboud University, 2020.

GDPR, a researcher would have to find data that fall outside the scope of the GDPR. One possible route to do this is by using anonymous data. Sometimes, training an AI-model on anonymised data is a viable option, but sometimes anonymisation of personal data leads to too much utility loss. The holy grail is then to find a surrogate to personal data that has the same utility yet does not qualify as personal data. One possibility – at least in most Member States – is to use data of deceased people (see section 4 below). Another one is to use so-called synthetic data, which are fake data that resemble real personal data. (see section 5 below).

4 A surrogate to personal data: Data of deceased people

Recital 27 explicitly states that the GDPR does not apply to the personal data of deceased persons, but that Member States may provide for rules regarding the processing of personal data of deceased persons. Approximately two-thirds of EU Member States have not chosen to do so.[38] For example, The Netherlands and Sweden have not created any provisions for data of deceased individuals; however, in Denmark, there is 10 years of protection after moment of death (§ 2(5) Danish Data Protection Act[39]). This means that in many Member States, data of deceased individuals could be a legal loophole and a window of opportunity for certain types of research. There are, however, several caveats to take into account.

Firstly, data of deceased people only fall outside the scope of the GDPR if they do not relate to any living individual.[40] A post on a social network containing information about both a deceased and a living individual would still qualify as personal data in the meaning of the GDPR. Certain types of data, such as genetic data, almost always also relate to living people even if they are primarily related to a deceased individual.

[38] David Erdos, Dead ringers? Legal persons and the deceased in European data protection law, 40 Computer Law & Security Review 40, 2021. See for an overview, for example: https://www.twobirds.com/en/in-focus/general-data-protection-regulation/gdpr-tracker/deceased-persons (last accessed 10 December 2021).

[39] Act No. 502 of 23 May 2018, published in the Law Gazette on 24 May 2018, at: https://www.datatilsynet.dk/media/7753/danish-data-protection-act.pdf.

[40] Iñigo de Miguel Beriain, Aliuska Duardo-Sánchez, José Castillo Parrilla, What Can We Do with the Data of Deceased People? A Normative Proposal, 29 European Review of Private Law, Issue 5 (2021), pp. 785–806.

Secondly, certain types of research need to be approved by a research ethics body. It should be noted that data protection and research ethics depart from different legal rationales: the former is strongly connected to informational self-determination, while the latter connects more to human dignity. National guidelines and regulations on research ethics often differ quite substantially from one country to another. However, some widely recognised international codes exist. One of the most important ones is the World Medical Association's Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects.[41] The first paragraph of its Preamble, the Helsinki Declaration, says that it applies to 'medical research involving human subjects, including research on identifiable human material and data'. Medical research involving human participants, genetic data or human tissue are classical types of research that are required to undergo ethical review in almost any country. Research from other domains, such as humanities[42] or natural sciences,[43] can sometimes also be required to undergo ethical review. In Sweden, the *Act concerning the Ethical Review of Research Involving Humans*[44] makes a link in 3§(1) to sensitive personal data, as defined in Article 9 of the GDPR: any research processing such data has to apply for ethical approval. It should, however, be underlined that despite the link to the GDPR research, ethics assessments follow their own logic that has to be clearly distinguished from an assessment of data protection compliance. Research ethical assessments often, for example in the aforementioned Swedish Act and the Helsinki declaration, have two main elements: independent ethical oversight that balances the scientific value against the privacy, health and safety it may entail for involved human participants' risks (and where priority is given to the latter) and informed consent. From a research ethics perspective, obtaining informed consent from study participants, unless they are deceased, is almost always necessary. This should be contrasted with the data protection law, where consent is only one

[41] World Medical Association's Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, adopted by the 18th WMA General Assembly, Helsinki, Finland, June 1964. Latest amendment on the 64th WMA General Assembly, Fortaleza, Brazil, October 2013.

[42] Ulf Görman, *Lathund för etikprövning – Humanistiska och teologiska (HT) fakulteterna*, Lunds Universitet, 2017.

[43] Etikprövning – en översyn av reglerna om forskning och hälso- och sjukvård (SOU 2017:104).

[44] Lag om etikprövning av forskning som avser människor, SFS 2003:460.

of several legal grounds for processing (Article 6(1) of the GDPR) and an enormous amount of processing happens on other grounds, without consent, such as the public or legitimate interest grounds. Thus, the 'freely-given, specific, informed and unambiguous' consent in data protection law differs 'conceptually and operationally'[45] from the informed consent that research ethics requires, the former rooted in information self-determination and the latter in human dignity. This distinction also explains why research ethics codes often do not exclude data relating to deceased people in the same way as the GDPR. For example, the use of biological material of a deceased human being might encroach on post-mortem human dignity, but informational self-determination is no longer applicable if an individual is not alive.

Thirdly, the question is if the right to private life in Article 8 of the European Convention on Human Rights (ECHR) could cause problems for the use of data relating to deceased people. This is rather unlikely. Despite the fact that the Strasbourg Court does not fully exclude that Article 8 can be applicable to deceased individuals,[46] the main focus is clearly on living people.

Finally, one should always take into consideration if data relating to deceased people are protected by some other rights of others, such as copyright, database rights or other intellectual property rights. This could, for example, be relevant with regard to pictures of deceased people that qualify as copyright protected works.

In summary, deceased people's data could be a good alternative for researchers who want to escape the scope of the GDPR, as long as the data does not relate to other living beings, and potentially applicable national data protection legislation about deceased people, research ethics and intellectual property laws are taken into consideration. Krutzinni and Floridi[47] have proposed that the use of medical data of deceased people should be facilitated by creating a dedicated medical code for posthumous data donation (PMDD) that enables individuals to decide how their medical data could be used after their death, in a manner akin to

[45] EDPS, Opinion on scientific research (no 2), 2.

[46] European Court of Human Rights (ECtHR), *M. L. v Slovakia* (Application no. 34159/17), 14 October 2021.

[47] Jenny Krutzinna & Luciano Floridi, Ethical Medical Data Donation: A Pressing Issue, in Jenny Krutzinna & Luciano Floridi, *The ethics of medical data donation*, Springer Nature, 2019, 1–6.

how one decides 'to donate blood, organs or tissue'.[48] While this proposal for a dedicated PMDD code has not been adopted by any legislator yet, it seems to fit in well with the spirit of the aforementioned DGA, and it is not difficult to see how data altruism could be stretched out to apply to this kind of posthumous data altruism too. While Krutzinna, Taddeo and Floridi[49] consider it ethically preferable to begin with creating a framework for posthumous data donation and later possibly extend to medical data donation by living individuals and corporations, the EU legislator seems to work from the other direction by introducing data intermediation services and data altruism in the proposed DGA.

5 A surrogate to personal data: Synthetic data (a particular type of anonymised data)

The material scope of the GDPR, as discussed above (in section 2), is very broad because of the enormous amount of data that fall under the definition of personal data, as defined in Article 4(1) of the GDPR.[50] This means that many researchers face GDPR questions, unless they find a way to escape its scope. Using data of deceased people as a way to escape the scope of the GDPR is only a minor fringe phenomenon compared to the most classical way to do so, namely by using non-personal data or by anonymising data.

Apart from the fact that national complementary provisions can make data protection extend to data of deceased people, such data also have practical limitations. Not all research can be based on data relating to deceased people. For example, in order to create AI models that capture contemporary phenomena, such as symptoms caused by the latest variety of the Covid-virus, outdated data of deceased people will not do. If data of deceased people will not help a researcher, it might be time to look at the more conventional road: anonymisation.

[48] Krutzinna & Floridi (no 47), 2.

[49] Jenny Krutzinna, Mariarosaria Taddeo, and Luciano Floridi, Enabling Posthumous Medical Data Donation: A Plea for the Ethical Utilisation of Personal Health Data, in Jenny Krutzinna & Luciano Floridi, *The ethics of medical data donation*, Springer Nature, 2019, 163–180.

[50] GDPR 2016/679 (n. 3).

In order to anonymise personal data, their link to any 'identified or identifiable natural person' (Article 4(1) of the GDPR[51]) needs to be severed in a way that cannot be reversed with 'all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly' (Recital 26 GDPR). This is not an easy feat. Removing a name or personal identification number is hardly ever enough to anonymise the data properly because the remaining data in a profile might be rich enough to single out a particular individual. For example, a data profile that refers to a female academic with Dutch origin working in public law at Uppsala University singles me out, despite the fact that it does not contain my name or some other unique identifier. Re-identification is often facilitated by the combination of different data sets or by using novel data techniques. This means that seemingly anonymised data, in practice, often actually should be qualified as pseudonymised data, defined in Article 4(5) of the GDPR as data that can 'no longer be attributed to a specific data subject without the use of additional information'. Pseudonymised data are a particular type of personal data, and thus still fall in the scope of the GDPR. This means that often, in order to realise true anonymisation[52] in the GDPR, quite a substantial amount of information should be removed, which can lead to a loss of utility. This is what is commonly known as the privacy-utility trade-off: by removing information, data might become anonymised, but this is of little avail if it disfigures the data to such an extent that they are no longer useful for the intended research. The holy grail of anonymisation is thus to find techniques that prevent re-identification while preserving data utility. During the last few years, synthetic data[53] have been proposed as potentially being this holy grail of anonymisation.[54] The basic idea is that instead of removing data, an AI model is trained on real data to generate fake data with the same statistical properties. An example would be to create a generative model that creates convincingly realistic portrait pictures of non-existing peo-

[51] GDPR 2016/679 (n. 3).

[52] Even though written when the GDPR was not in force yet, many of the arguments are still applicable: Working Party 29, *Opinion 05/2014 on Anonymisation Techniques*, 2014.

[53] Luke Rodriguez & Bill Howe, In Defense of Synthetic Data, arXiv:1905.01351, 2020; Anjana Ahuja, The promise of synthetic data, Financial Times, 2020; Laboratory for Information and Decision Systems, The real promise of synthetic data, MIT News, 2020.

[54] Steven M. Bellovin, Preetam K. Dutta, Nathan Reitinger, Privacy and Synthetic Datasets, 22 Stanford Technology Law Review, Issue 1, 2019, 1–51.

ple[55] and to use these simulated data as a basis to train an AI model. The question is, of course, if training a model on synthetic data will generate sufficiently good results in comparison to using real data. Some authors[56] have claimed that this might not be the case and that synthetic data of good quality might still be traceable to identifiable or identified individuals, and that the privacy-utility trade-off is not truly resolved by using synthetic data. Despite the drawbacks, using synthetic data might be a viable option for at least some types of relatively simple data (for example, a portrait photo might be more easily simulated than a brain scan) with little outliers (for example, if a data set of portrait photos contains 100,000 human faces and 5 cat faces, the cat pictures in the synthetic data set will probably be much closer to the original pictures than the human ones, simply because there has not been enough material to generalise).

6 Conclusions: There is no one size that fits all

What to tell a researcher who needs data to build an AI-model but fears that the GDPR-requirements will create a burden that is too large and too demanding? The first message is a comforting one. The GDPR has a broad understanding of scientific research, and it has a rather generous research exception in Article 89 of the GDPR. Nevertheless, GDPR compliance is more than just ticking a few boxes and will often require some substantive thinking about data protection risks and balancing of different interests. The EU legislator partially helps researchers in the proposed DGA, by creating infrastructures that will help match potential data sharers and data users. Data intermediation services and data altruism organisations are thus likely to increase data access to data that are protected by rights of others (data protection or intellectual property rights). However, it should be underlined that the GDPR has primacy over the DGA and that the improvements are mostly in terms of data availability and infrastructure. The potential burden of compliance with GDPR requirements is not altered by the proposed DGA. For researchers that want to use data and not be burdened by the GDPR, I discuss two alternatives. Firstly, one could consider using data of deceased individuals, in as far as they are not connected to any other living individuals.

[55] See https://thispersondoesnotexist.com/ (last accessed 10 December 2021).

[56] Theresa Stadler, Bristena Oprisanu & Carmela Troncoso, Synthetic Data – Anonymisation Groundhog Day, arXiv:2011.07018, 2022.

Here, also other legislative instruments should be taken into account that could potentially limit the ways in which such data may be used: national data protection provisions on data of deceased individuals, research ethics codes and intellectual property laws. Secondly, one could consider using anonymised data. In those cases where traditional anonymisation methods degrade the utility of the data too much, the use of synthetic data could be an option.

In summary, the researcher suffering from GDPR-anxiety, who is looking for personal data or surrogates with a similar level of utility, in principle, has a *smörgåsbord* of options to pick from. However, which type of data will be helpful in a particular research project is a highly contextual question – in finding the right type of research data, there is no *one size fits all*.