Data Altruism by Default: An Alternative to Consent for Personal Data Processing in Machine Learning

Tea Mustać*
* Spirit Legal Rechtsanwaltsgesellschaft mbH, Neumarkt 16-18, 04109 Leipzig, Germany
tea.mustac@spiritlegal.com

Abstract. This paper is not about data altruism in the lexicological or grammatical sense of the word. It is also not about data altruism in the Data Governance Act sense of the word. This article is about carving out an exception for certain types of data processing conducted for the development of AI systems. All in the attempt to allow the development of high-quality AI systems while striking an appropriate balance between the developers of such systems and the data subjects providing the data for their development. This paper aims to assist in confronting these challenges by critically examining the monetization of personal data obtained through dubious and largely unlawful practices. Labelled provisionally as *data altruism,* such invasive data practices deny the users any real choice. On the other hand, *data altruism* proposed in the Data Governance Acts does little to resolve the existing issues as the provided alternative is overly limited in its scope. In contrast to this, the paper provocatively proposes a novel framework which would justify the necessary processing practices while imposing objective and enforceable requirements on those engaging in them. In light of this, the paper underscores the need for robust safeguards and highlights the importance of objective limitations to associated business models. The proposed version of data altruism is crucially different then both previously examined, it should, however, not be understood as a solution but rather as a provocative idea stimulating meaningful discussion at both the societal and political level.

Keywords: Data Altruism, Data Monetization, Artificial Intelligence, Transparency, Consent.

1 Introduction

Technology often confronts us with some complicated questions. In that sense, artificial intelligence ('AI') technology hardly brings anything novel. Except maybe the fact that superintelligent world takeovers have now exited the world of science fiction and have made their way into popular scientific literature. However, regardless of the admittedly tempting appeal of such questions, there are much more urgent problems in need of our attention, lest we want to have our futures determined for us. The way we handle these issues is bound to have a direct effect on any bigger philosophical issues we wish to discuss in the aftermath. One example of such an issue includes AI developers taking data for free and often through unlawful practices to then use this data for developing products and services to sell on the market. This, what seems to be, front-end *data altruism* (or stealing, depending on the point of

view) coupled with back-end monetization of that same data is one of the less obvious issues of AI system, however, it is deeply intwined with other more prominently discussed ones, such as explainability or transparency.

In this paper, we will focus on the question of whether we can establish an alternative in the law, which would make the data processing necessary for developing functioning AI systems lawful while still protecting the rights of data subjects. Turning this imposed *data altruism* into something resembling *real* data altruism. To do so we will try and reconcile the requirements of the existing regulations, GDPR in particular, with the large amounts of high-quality data necessary to train a well-functioning model. This is no easy task, which is also probably why it has gotten comparatively little attention in the current and mostly theoretical discussions. This paper will try to rectify that. To this extent, the focus of the paper will be on the currently prevailing and mostly unlawful data collection practices analyzed on the example of OpenAI's privacy policy. Then we will take a look at the data altruism framework proposed in the Data Governance Act (DGA) to demonstrate where it fails to remedy the existing issues. Finally, a novel strategy for dealing with the identified challenges referred to as Data Altruism by Default will be briefly presented. The proposed framework should not be understood as a *solution* waiting to be implemented into existing laws, however, it can hopefully instigate further research and discussion, as to how existing standards can be applied to the development and deployment of AI technologies. As well as, if they cannot, help provide an alternative solution and open it for societal and political discourse.

2 Transparency as the Prerequisite of Lawfulness

The idea of imposing transparency on actors in society is almost as old as law itself. It can be traced back to ancient Rome, when *Appius Claudius Caecus* composed the *formularies* of action necessary to start a civil trial to simplify citizen access to legal remedies. [1, 2] Over the ancient Greek ideals of transparent democratic process and governance [3], all the way to transparency in the modern sense of the word. Thus including, for instance, the rich jurisprudence on transparency in consumer contracts [4] as well as transparency concerning data processing. However, especially when it comes to new and complex technologies, complying with transparency appears to be shifting from honoring a fundamental legal principle to *selling* a kind of narrative.[5,6] A story told to the data subjects about what is done with their data, based on certain expectations about what a particular user can understand, but also what the user wants to hear. The mental exercise consists of saying just the '*right'* amount of the '*right'* things to make the user agree to the processing. And that is if the user is lucky, and the provider even tries to comply with the existing regulations. The strategies of many AI providers range from bad to even worse when it comes to respecting the rights of end users. Not to mention the rights of any person that has ever accessed the internet as their data is also used in the training of most AI models and there does not appear to be a thing anyone can do about it. Now leaving blatant disregard of existing laws and judicial interpretations aside, such as for instance for the conclusions of the European Court of Justice ('the Court') stating that the fact that someone clicked, liked or shared something online, does not mean the person also

made the associated personal data manifestly public and therefore free to use as the AI developers please.[7] So even when the AI developers and data controllers try to honor the requirements of transparency and engage in lawful data collection and processing, there are still many challenges they face. Paired with even more strategies for avoiding them. These will be briefly elaborated in the following Sections.

2.1 The Requirements of Transparency

The GDPR takes a broad notion of what it means to '*process data'*. This includes everything from collection all the way to anonymization and deletion. And the essential principles that have to be followed to engage in any processing operation lawfully is to do so transparently. Transparency itself is yet another complex principle consisting of multiple further requirements. We can start the discussion with one of them, namely that of providing concise, transparent, intelligible and easily accessible information, which we will juxtapose to the practice of making privacy policies and information provided difficult to find and even more difficult to understand. Taking, as an example, the privacy policy of OpenAI. [8]

The company offers various tools and services, with their privacy policy supposedly covering them all, without specific reference to any of them. And it gets worse. The policy is short and very vague, with a lot of general and nothing-saying bullet points such as:

> *"For example, you may have the right to: Access your Personal Information and information relating to how it is processed."* or

> *"As noted above, we may use Content you provide us to improve our Services, for example to train the models that power ChatGPT. See here for instructions on how you can opt out of our use of your Content to train our models."*

One major improvement on the side of OpenAI that has to be mentioned is that, somewhere towards the end of 2023, the company implemented a pop-up telling the users that "*Chat history may be reviewed and used to improve [their] services*", appearing as soon as one accesses the chat. Meaning, at least the user no longer has to actively seek out OpenAI's privacy policy and read it in order to find out that the input data is used to train AI models. Still, paragraphs like these hardly provide any meaningful information to the user as to which data is collected and processed for "*train[ing] the models that power ChatGPT*", which models that may be or how to exercise the stated right. Furthermore, and to just focus on the second paragraph in particular, the fact that the user has to actively opt out of any such processing is still irreconcilable with the standards imposed by the GDPR. Finally, should the users have any questions about the policy, they are instructed to click the link "contact support", in which case they will be redirected to a kind of forum website with app-specific FAQ.[1] The forum itself is a significant improvement, with much more detailed information on the processing operations and graphics showing how to

[1] See, for example, https://help.openai.com/en/articles/7730893-data-controls-faq (accessed on the 8 of January 2024).

(again) opt-out of certain per default conducted data processing operations. The one thing the user is, however, unable to do is actually contact human support to ask a specific question. One is always free to discuss the issues with their 'Help' chatbot, however.[2]

The second requirement of transparency is that the language of the provided information is clear, plain and simple. However, even if the language meets this condition, it still does not mean that the information explains the relevant processing operations in the same fashion. Taking again the example of OpenAI, there can be little argument as to whether the description of collected data as "*Network Activity Information, such as Content and how you interact with our Services*", uses simple language that a person can linguistically understand. However, whether the person would also consider that this entails everything from the moment they accessed the service, over the period they used it for, up to the particular textual input to the system, can be further discussed. At this point, explaining the full complexity of the processing operations is also to a larger extent irrelevant, as the user does not necessarily need to know how the sentences *prompted* to the system are *tokenized* or *embedded* into the *vectoral space*. It is also irrelevant to understand how particular data units are used to adjust the *weights* and *parameters* of the system. Oftentimes not even the developer of such systems can answer these questions with complete certainty. On the other hand, what the users should know and what is also easy to explain using simple language is that all their account information as well as textual inputs are used to gather feedback on the model's performance and improve it in the future. One could even argue that the requirement of using 'clear, plain and simple language' mandates avoiding *vectors*, *tokens* or *memorization* in the explanation. What this requirement is about is providing the user with a basic understanding about what is going on as well as why it is necessary (if it is in fact necessary).[3] And this can indeed be done.

One common defense against investing the effort of providing the necessary information in a transparent and easily understandable manner and prior to commencing the processing is so-called "information fatigue". (We could also add "legalese fatigue" [11], "privacy fatigue" [45] and "consent fatigue" [46, 47] to this, but these are all topics for another discussion.) Information fatigue is often used as an excuse for getting out of sharing any information with the data subjects to avoid them being overwhelmed. Conversely, one thing definitely not fatiguing anyone is the information in the OpenAI's privacy policy. For the simple reason that there is not

---

[2] See, How can I contact support?, OpenAI, https://help.openai.com/en/articles/6614161-how-can-i-contact-support (accessed on the 25 of December 2023).

[3] There is also ample evidence to claim that basic understanding is not sufficient and that data subjects have the right to receive both accessible and sufficiently complete information on, e.g. automated decision making, which would allow them to fully comprehend the processing operations and the effect these may have on them. See, for example, Ppinion of Advocate General Richard de la Tour delivered on 12th of September 2024, Dun & Bradstreet Austria Case C-203/22, para.76. However, as the argument being made in this paper is that not even basic levels of understanding are achieved, such stricter interpretations only serve to support the claim that the current framework is poorly designed to enable advanced technologies and support the associated data requirements.

much there. But is there such a thing as middle ground here? Furthermore, maybe in certain cases we would want the users to be overwhelmed? Maybe if they could see how much space it takes to describe what is being done with their data they would think twice before registering for the latest AI tool or buying the latest smart gadget? Still, scaring the users to prevent them from consenting to sketchy processing operations is probably not what we want. What we do want is for them to understand the data flows and processing operations, understand why these are necessary (if they are in fact necessary) to access a particular service, and to consent to this. Not because they could not be bothered to read the notice but because they understood and acknowledged it. This may sound like utopia. Wishful fantasies as to how life may look like in Neverland. And yet there are solid examples for how this can be done. Whether you want to draw it up [12] or make a video out of it [13], there have been examples of successful stories making the information digestible and simple. [45] And then it can only ever be left to the users to decide if they will engage with the information. However, now comes the plot twist. Previous success stories most often concerned (arguably) very simple processing operations. Or at least much simpler than whatever is going on inside large language or diffusion models. Not to mention they involved companies who had direct contact to the data subject.

Nonetheless, when we are talking about enterprises such as OpenAI, currently offering their products to about 180.5 million people across the world. [9] An enterprise that now also charges for certain services it provides, [10] blatant disregard of regulations, including the most basic transparency requirements, should be intolerable. And yet we tolerate it. We tolerate that the users are not explained which data is being collected or how their data is processed, but instead have to actively seek out and collect this information piece by piece. And not only that but the user also has to decipher the information collected from privacy policies linking to forums linking to FAQ to find out what is going on with the data. And although understanding the inner functioning of an LLM model might require a computer science degree, this is still no excuse to not even provide the most quintessential information. Nor does it justify having to actively object to non-essential processing operations. At least not under the existing legal framework. The key here is providing the essential processing information to the users and not giving them a vague bullet point *cheat sheet* with the title Privacy Policy.

2.2 **Transparency, AI Literacy and Trustworthy AI**

When it comes to AI systems, transparency is often mentioned as one of the key requirements not only for collecting training data, but also for achieving one of the most elusive concepts in the AI universe. That of developing ‘trustworthy’ AI. This term, also proclaimed as one of the goals of the AI act [20], can mean everything and nothing. Especially, due to the emotional electricity of the term. As trust, usually associated with the feeling we have towards other human beings, maybe our pets and if we are lucky ourselves, is something extremely difficult to define or even recognize. Especially as, now, it is associated with machines and algorithms. This feeling can hardly ever be standardized, as there is no definite rule as to when it develops or when it dissipates, and the answer to such questions vary greatly among people and cultures. When considering all that, using this concept as a goal of

regulating complex technologies seems futile at best. However, as discussing the deficiencies of this concept is almost as futile as the concept itself, we can ponder on what a trustworthy system would look like.

In general, and while affinity for technology can hardly be rewired in people, we can try and work on other factors determining the likelihood of a person trusting an AI system. Or at least not distrusting it. And these are prior knowledge and understanding of the technology in general, as well as regarding the particular system in question. What we can also work on the factors influencing the likeliness of the users accepting the technology in line with the Technology Acceptance Model [21], which are perceived usefulness of a technology and user experience while interacting with it. Needless to say, that the system should also be accurate and secure to the greatest possible extent. At this point, however, we will focus on enhancing knowledge and understanding for increasing the trust people have in AI systems in general, especially by working on AI literacy, which is now also an obligation under the AI Act. Teaching people that they are interacting with something highly mathematical that works by calculating probabilities and extracting patterns can be a good place to start. However, explaining this also goes way beyond traditional requirements for transparency. As well as, in most cases, well beyond that what most (small and medium sized) developers of AI systems are in a position to do. Nonetheless, having short videos explaining which technology a particular system uses and how this technology works is a solid starting point. Furthermore, educating the larger public is also in the interest of AI developers, so that the users build '*trust'* in the technology in general and their relevant system in particular. Therefore, it is also not uncommon for some of the biggest players, who have recognized this, to already pursue this strategy. For instance, Google and DeepMind have been offering courses and having free explanatory videos on their technologies for years.[4] Linking any such materials in the privacy notice, will, in any event, help shorten the notice while making sure the information is easily and readily made available for any user interested in accessing it as well as help fulfil the requirements of AI literacy.

Finally, transparency is also subject to well-developed standards under the GDPR and the Directive 93/13/EEC. As we have analyzed some of the requirements of transparency in the previous Section, we will refrain from doing it here again. We will rather simply emphasize that many of these obligations come regardless of the AI literacy discussion and before there is even an AI system to be literate about. For instance, all data subjects, whose data was collected to train an AI system should be informed of the fact prior to the collection. This requirement is extremely difficult to comply with, as many of these systems were trained on data scraped from the internet, regardless of whether or not a person is or was ever interacting with a particular AI developer. Secondly, even if the eveloper used data scraped by a separate entity, from the moment they start using it, they still become a data controller in their own right. This in turn again requires them to inform the data subjects that their data was obtained from a third party and used to train the systems. One popular defense against

[4] For example, Google offers dozens of courses on the popular learning platform Coursera. As well as dozens of videos on YouTube, a lot of which have an educational component. See, for instance '5 essentials to know about generative AI from Google', https://youtu.be/unPKJJjQP0A?si=L9bSfubda4eCPLlC (accessed on the 25 of December 2023).

complying with the requirement from Article 14 of the GDPR, mandating that data processing information is also provided to data subjects, whose data was collected indirectly, is the claim that providing this information would involve disproportionate effort. Hence, no information needs to be provided to anyone. Now while this argument is understandable, if this is in fact the case this needs to be determined by European courts and relevant authorities. Especially since it diverges from some of the earlier decisions, which set the bar extremely high and mandated that all data subjects need to be individually notified of the processing even e.g. when there were as many as 5.7 million of them and the data collection involved *scraping*.[22, 23] Also, if relying on the Article 14(5) GDPR exception is a viable route to take, this needs to be analyzed in respect to what it means for the right to privacy in the age of AI systems. One answer imposing itself is that, when it comes to the development of these systems, the data subjects become mere spectators having no control over their data but rather being forced to altruistically give it away by the mere act of having accessed the internet. In light of the meaning of the word altruism, as the "*willingness to do things that bring advantages to others, even if it results in disadvantage for yourself*"[24], it can, however, never be forcefully imposed. Tacitly choosing this as our policy is already problematic with the problem emphasized on the back end, where this altruistically provided data is being monetized. This holds true regardless of the benefits AI may eventually and in certain aspects bring to humanity. This can only ever be a choice society makes and not a practice we unintentionally slide into.

All these issues as well as the general lack of high-quality data necessary to develop high-quality AI and the non-existent data sharing economy in the EU are not breaking news. Quite to the contrary, these issues have been discussed for decades and have even already reached the point of a regulatory intervention in the form of the Data Governance Act. However, this Act does not solve the problem. It, to the contrary, arguably creates even more unclarity in the field. We will briefly go over why that is the case in the next Chapter.

3 An Alternative to the GDPR?

The Data Governance Act officially entered into force in 2022 with its provisions being applicable from September 2023. In general, DGA aimed to set the conditions enhancing data sharing practices in Europe and enabling the development of common European data spaces. The EU despite its stringent data protection regulations is generally aware of the fact that data is an essential resource in today's economy.[48] And despite of the increasing amount of data being generated in the digital economy, its usefulness remains limited due to the non-existing data sharing framework in the EU. [48] The DGA is the first regulatory instrument aiming to rectify this,[5] especially by introducing *data altruism*, which as opposed to the forceful data collection presented in the previous section means a voluntary sharing of data for objectives of general interest. However, there are a lot of limitations associated with this data economy *facilitator*. We will briefly go over them as well as explain why these limitations make data altruism so construed, rather pointless.

[5] This is also true for the Data Act, but we will not go into that in this paper.

3.1 Data Altruism in the Data Governance Act

Data altruism is envisioned as voluntary sharing of both personal data by data subjects as well as non-personal data by data holders, without seeking or receiving a reward going beyond compensating their efforts, for objectives of general interest. When reading this definition, one cannot help but wonder what value exactly the related provisions aim to generate.

Firstly, this voluntary sharing is based on consent for all shared personal data and as the Data Governance Act refers to the GDPR for the definition of consent we can go straight back to the previous Chapter and re-read all the issues associated with collecting lawful, informed and meaningful consent. And we are not even going to get tangled in the persisting implications that withdrawing consent may have, both GDPR and DGA-wise. Because despite the fact that some authors see a critical difference between the two consents, [49] the fact that one can only be given for specific, 'altruistic' purposes and might at some point be harmonized by a data altruism consent form template, does not merit a separate status from the GDPR consent. One valid point of criticism that has been raised deals with the question of whether such a form, once it becomes available, has to be used in order to make the consent valid. [49] Still, regardless of whether using this consent form will be mandatory or not, the only thing such a form could ever achieve is raising the trust that the actor is actually a public body engaging in practices of general value, as well as possibly diminish the likelihood of malevolent actors exploiting the regime. Conversely, this would never mean that the GDPR requirements do not apply.

Secondly, data altruism in general can (meaning does not have to) be provided by member states and it can only be exercised for objectives of general interest that are provided in national law. The notion of 'general interests' is clarified using examples such as healthcare, combating climate change and improving mobility. In this context, it remains unclear if 'general interest' from the Data Governance Act is the same as 'public interest' in the GDPR. Probably it is not as processing personal data to pursue activities of public interest is already allowed under the GDPR, so equating the two would make data altruism as such superfluous. Or, at best, it would make determining the legal basis for such processing operations very complicated. [50] Moreover, clarifying the notion in a way that distinguishes it from public interest and specifies its content is also necessary as the DGA should be a directly implementable regulation, which the data subjects can directly call upon to exercise their rights. Without clarity on the interests that justify data altruism no such direct invocation is possible. [50] This, together with the conditions for an organization to register as a 'data altruism organization' in the first place, significantly limits the number of entities as well as situations which could ever rely on such an exception.

Thirdly, why data subjects would actively seek out these organizations and volunteer their data to them or how they would even find out about the existence of such organizations in the first place, remains elusive at best. Even if people would be generally willing to give away their data for the 'greater good', without a direct incentive in terms of a public service they can receive for sharing their data, it is hard to imagine individuals proactively seeking out these organizations and providing them with data without having at least an idea of what the altruistic organizations is going to do with that data. Of course, public awareness raising campaigns might help

stimulate such practices, but it remains highly unlikely that this type of data altruism will ever generate the necessary amounts of data for training ML models, and even if it does this will be restricted to public bodies developing AI systems of general interest, where data sharing in general is not as big of an issue and we already have examples of successfully conducting such projects.[51]

3.2 **The Essence of Privacy**

One final question we will consider in this section, especially important when we are discussing highly complex processing operations and novel concepts such as data altruism is: are there any limits to what one can consent to? This question is particularly important to determine whether there is data that is completely of limits. The question of the existence of limits to consent in general will depend greatly on what one is consenting to. And as the right to data protection is considered a fundamental human right under the European Convention on Human Rights (ECHR) [25, 26] as well as under the Charter of Fundamental Rights of the European Union, [27] the question cannot be overlooked in this context either.

Most countries today are paternalistic towards their citizens. For instance, in the European Union, we can never (lawfully) consent to end our own lives, no matter how hard the circumstances. Or, on a less dramatic note, to drive our car without a buckled safety belt. At least not without risking getting a fine. Thus, the question of whether one can consent to being a completely open book towards any organization and especially towards commercial entities, whose business model is based on exploiting this fact, cannot be simply shrugged off. For instance, Bock and Engeler have argued already in 2016[28], that the users can never consent to intrusions into their privacy as a fundamental right and the article makes some critical points. For instance, the authors state that data protection needs to be established independently of the right to privacy, as the two are often intermingled and confused. Privacy and data protection developed separately and have to be understood as such. Therefore, while the right to privacy may protect us from a curious neighbor peeking through our window, our right to data protection aims to protect processed data units about us. Furthermore, while privacy can be considered a rather subjective concept, data protection can potentially be conceived more objectively. Thereby also offering a unique opportunity to delimit it objectively by putting certain data "*extra commercium*" [29]. To a certain extent the GDPR already does that, as it puts certain special categories of data (e.g. Article 9) on a much higher level of protection than other. Consequently, one would simply need to consider (or reconsider) if there are any *extra* special categories, quantities or purposes of data and data processing, that should be completely out of reach for commercial entities. Finally, this would also not be completely unprecedented, as for instance the newly adopted Directive on credit agreements for consumers does precisely that by stating in Article 18(3), that data collected from social media can never be used for credit scoring.[31]

As mentioned, this question gets all the more important when new and complex technologies, such as generative AI, and very novel ideas, such as data altruism, are considered. Does the sensitivity of the data, paired with aggressive, manipulative processing practices and the complexity of the underlying processing operations, merit special frameworks? Does ML training, at least in some cases, merit carving out

another exception in the GDPR for training AI systems? [42] If yes, are there any categories of data that should be specifically excluded from the scheme? All valid and important questions currently flying under the radar, as we are all busy trying to wrap our minds around what is even going on in the first place. Finally, while we are busy with all the *wrapping*, another important thing happens. The users are sliding into accepting whatever bubble pops up on their screen, if one even pops up that is, and continue to voluntarily provide all their data to anyone who cares enough to ask.[6]

4 Data Altruism 2.0

4.1 The Necessity for a Novel Legal Basis for Machine Learning

Considering the highlighted difficulties of collecting effective consent just for the development or use of complex AI systems, as well as the very broad notion of what is considered personal data in the EU and what rights the data subjects have with regard to that data, many of which are impossible to honor for many AI system developers, the situation should be pretty straight forward. Data was collected and/or processed unlawfully and can be extracted from many of the models [41] *ergo* models are unlawful (or even personal data themselves) [41] and should be deleted. Yet we have not even reached the point of demanding the deletion of ClearviewAI's systems[7] and when it comes to some of the more generally useful and popular commercial AI systems, such as ChatGPT, most are still busy trying to pretend that they are lawful despite extensive research claiming otherwise.[8]

---

[6] It would be unfair to claim that this issue is completely neglected as there are people researching the matter and raising awareness. To this extent, see 'The Biggest Lie on the Internet' which is a website part of the Understanding "the Biggest Lie on the Internet" Project. Website is available here: https://www.biggestlieonline.com/ (accessed on the 5 of January 2023).

[7] A summary of the current efforts of European data protection authorities can be found here: Challenge against Clearview AI in Europe, Privacy International, https://privacyinternational.org/legal-action/challenge-against-clearview-ai-europe (accessed on the 20 of September 2024).

[8] For example, recently the Hamburg Commissioner for Data protection and freedom of information has published their Discussion Paper: Large Language Models and Personal Data, in which they claim that LLMs do not store any personal data and therefore the data subjects rights are not relevant. See here: https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Discussion_Paper_Hamburg_DPA_KI_Models.pdf. This is in stark contrast with both the current research efforts which continuously proves how easily the training data can be extracted from these models as well as the stance most data protection authorities take on the matter. See, for example: Li, H., et al. Privacy in large language models: Attacks, defenses and future directions. (2023) a*rXiv preprint arXiv:2310.10383*; Pan, X., et al. Privacy risks of general-purpose language models. (2020, May). In *2020 IEEE Symposium on Security and Privacy (SP)* (pp. 1314-1331). IEEE; Zhang, D., et al. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. (2024). *AI and Ethics*, 1-10; Lareo, X., Large language models (LLM), European Data Protection Supervisor (accessed on the 19 of September 2024).

They are not lawful, and they will never be lawful, unless we carve out an exception making the necessary processing operations lawful, regardless of the GDPR, and impose the conditions for the applicability of such an exception.[9] And even though such an exception is somewhat unprecedented in the GDPR context,[10] a parallel could be made with the text and data mining exception introduced in the Directive on Copyright in the Digital Single Market. Copyright has a somewhat longer history than data protection, which can be traced down all the way back to 6th century and the Irish legend of St. Columba who first copied a book without permission and caused a war as well as the establishment of the first rule of copyright: "*To every cow its calf to every book its copy.*"[11] Roughly translated as who owns the book also has the right to its copies. Understandable. Still, as centuries went by, novel technologies raised novel and slightly more complex questions. From the rise of the printing press, the invention of cameras and scanners, all the way to machine learning, the situation was seldom straightforward, and existing concepts often had to be reconsidered in an open dialogue between the authors and the regulators trying to also take into account the interests of the wider public. The last time these negotiations took place they resulted in the Directive on Copyright in the Digital Single Market (CDSM) and the official introduction of the text and data mining exception, which was meant to bring some clarity in terms of how and when copyright protected content can be used for training machine learning systems.

In the CDSM, the text and data mining exception is mentioned twice. Once as an exception for the purposes of scientific research (Article 3), which was also the primary reason for introducing the exception in the first place.[43] And a second time for all other purposes, meaning also commercial (Article 4), which is then associated with more stringent conditions for relying on the exception. In this context, 'text and data mining' should be understood as any automated analytical technique aimed at analysing text and data, including sounds and images, (Recital 8 of the CDSM) in digital form in order to generate information which includes but is not limited to patterns, trends and correlations (Article 2(2) of the CDSM). According to Recital 8,

---

[9] Similar studies have already been conducted regarding the use of sensitive data for detecting and preventing algorithmic bias. See, for example, van Bekkum, M. and Zuiderveen Borgesius, F., 'Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception?,' Computer Law & Security Review, Volume 48, 2023, https://doi.org/10.1016/j.clsr.2022.105770.

[10] Only somewhat, as Article 85 of the GDPR recognizes processing personal data for journalistic, academic, artistic or literary expression purposes as meriting exemptions and derogations from Chapter II to VII as well as Chapter IX, if such measures are necessary to reconcile the right of data protection with the freedom of expression and information. Furthermore, Article 89 similarly proposes that safeguards and derogations be implemented in respect to some data subject right when processing personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.

[11] To read more on the legend, see Connolly, S., The Story of St. Columba and the Book of Kells, The Catholic World Report, June 9, 2023, https://www.catholicworldreport.com/2023/06/09/the-story-of-st-columba-and-the-book-of-kells/ (accessed on the 20 of September 2024). Masnick, M., The Very First Copyright Trial, In 6th Century Ireland, Sounds Really Familiar, techdirty, August 20, 2009, https://www.techdirt.com/2009/08/20/the-very-first-copyright-trial-in-6th-century-ireland-sounds-really-familiar/ (accessed on the 20 of September 2024).

such an exception was necessary as "*text and data mining can involve acts protected by copyright, ... which occur for example when the data are normalised in the process of text and data mining."* And this is important because *"[w]here no exception or limitation applies, an authorisation to undertake such acts is required from rightholders*." We can learn a lot about the lawmakers' intentions from reading the Recitals of the Directive.

1. The exception was introduced as a means of achieving balance and protecting open research.
2. The exception was necessary as TDM often involves copyright relevant reproductions, meaning permanent reproductions of copyright protected works.
3. Due to the initial intention of protecting research it is justified that there are no reservations, which can be made to the exception.
4. Even though the initial intention was to protect open research, it was necessary to enable text and data mining also for commercial actors and systems, but only under further conditions (Recital 18 of the CDSM).

The two main conditions for relying on the TDM exception for commercial purposes include that the work is accessed lawfully and the rightsholder has not made any reservations preventing TDM. Such reservations should be made "*by the use of machine-readable means, including metadata and terms and conditions of a website or a service*." However, in some cases it may also be appropriate to reserve the rights otherwise such as through contractual agreements and even unilateral statements. From this Recital it is clear that:

1. Commercial actors have a recognized right to engage in text and data mining.
2. This exception only applies, if the access to the works in question was lawful.
3. The access is always unlawful, where the rightsholder made any reservations to such commercial mining.

This exception, in line with the copyright tradition, has to be interpreted narrowly and always to the benefit of achieving a fair balance of interests, meaning also that the end product can never be used in ways that conflicts with normal exploitation of the work or unreasonably prejudices the legitimate interests of the authors.[12] Similarly to how the copyright and IT community came to a conclusion that a novel solution was necessary to enable use of high-quality data for machine learning, we can also imagine doing the same for using personal data for the purposes of machine learning. Despite the fact that there are still many discussions regarding the text and data mining exception, it has helped clarify the legal situation and simplified certain types of machine learning, at least for non-profit entities. Therefore, we can reasonably assume that a similar exception applied to personal data would also enhance legal clarity and legal certainty with which market actors can act in the market, as well as allow us to set up objective rules and requirements as to how, when and under what conditions the exception would apply. This would ultimately also help enforcement as

[12] See Article 9 The Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979).

such clear rules would allow better supervision and faster action upon noticed violations.

4.2 An Argument for Data Altruism by Default

After the previous section some are probably still unconvinced. Why do we need an exception and why does it have to apply by default? To these and similar questions, the following arguments can be offered.

Firstly, the processing operations which would be tolerated under this exception albeit apparently standing in stark contrast to data protection principles such as data minimization or storage limitation, actually do not concern these principles at all. Namely, despite the fact that many of the used data qualifies as personal due to the ever-persisting theoretical possibility of data subjects being identified by someone at some point, this would be neither the purpose nor a wanted side-effect of the processing operations which could benefit from the exception. There can be no doubt that skillful and malevolent actors can be very creative in finding ever new ways to identify data subjects regardless of the intentions of the data controller.[13] However, completely preventing or imposing impossible conditions onto the developers is not a justifiable consequence of that fact. Same as we do not blame a bank taking all necessary precautionary measures for being robbed by a skilled criminal, but instead demand them to have securities and insurances in place to be able to compensate any damages occurring in such cases, we cannot blame a diligent and cautious data controller when a skilled hacker manages to break their servers and identify individuals whom the data belongs to. Especially, since their original processing purpose had nothing to do with the identity of the individuals whose data is contained in the data sets and has everything to do with extracting statistics and recognizing patterns in those data sets. Furthermore, applying such stringent criteria would already prevent almost any type of ML processing, due to the simple fact that we lack high-quality and relevant data in general and compliance is complicated.

Secondly, as for the '*happening by default part'*, this is necessary as our legal issues begin far before we ever reach data subject. They begin at (often indirect) data collection and the issue of having a legal basis for processing the collected data. To step away from LLMs and the data their providers scraped of the Internet and continue scraping from their own social media sites to use for training for a moment,[14]

---

[13] Not to say that this cannot also happen unintentionally. For example, see the incident that happened at Tesla last year, where the workers not only had access to individual user recording but also shared the recordings and commented on them in their chat-rooms: Stecklow, S., Cunningham, W., and Jin, H., 'Tesla workers shared sensitive images recorded by customer cars', Reuters, 6 of April 2023, https://www.reuters.com/technology/tesla-workers-shared-sensitive-images-recorded-by-customer-cars-2023-04-06/ (accessed on the 19 of September 2024).

[14] For example, X has recently stopped using European user data to train its AI, bit only after already publishing the next version of their model, which was developed using user data and without even notifying the data subjects. More on the story can be found here: Lomas, N., 'Elon Musk's X targeted with nine privacy complaints after grabbing EU users' data for training Grok', TechCrunch, August 11, 2024, https://techcrunch.com/2024/08/11/elon-musks-x-targeted-with-eight-privacy-complaints-aft

we can imagine a vehicle collecting visual footage to be used for developing autonomous vehicles. Such a vehicle would somehow have to inform all the by passers that it is recording its surroundings, and therefore processing their data, as well as of their GDPR rights and how to exercise them. This applies even when we rely on legitimate interests for the processing, which most AI developers do, as even then the data subjects still have the right to object to the processing or restrict it and one does not have to be a computer scientist to figure out that would be impossible for a vehicle driving around. Not to mention that we would also need to conduct a thorough and detailed Legitimate Impact Assessment, which is something nobody is paying much attention to nowadays. [44] Finally, regardless of the legal basis and the personal character of the collected footage, the collected recordings would not be processed in a way that allows identification in the first place, whereas quite absurdly the individuals would have to be identified in those recordings if we were to honor their objection or even just confirm they were in fact on the recording. Compared to this, notifying the users of your own social network that you are using their posts and interactions to gather training material for your LLM seems comparatively easy, and we are currently not even getting that right.[15]

Thirdly, as in our example the recordings are only made for the purposes of gathering training data and developing functioning systems for object recognition and semantic segmentation, once the statistical patterns are extracted from the gathered data, that data can also be promptly deleted. And the probability to then extract any type of individual personal data from the models aimed at predicting the probabilities of next movements of traffic participants is significantly lower, not only due to the skill and effort that would be necessary to conduct such an attack but also due to the significantly decreased benefit the extracted data could ever bring to the attacker, who would hardly be able to find out anything other than confirming that data of a particular individual was in fact in the training set. Therefore, due to the apparent irreconcilability of the mere data collection with the GDPR, as well as its in many aspects unproportional requirements, it is essential that there is an exception for such processing and that the exception applies per default if certain conditions are met. In the final section we can briefly outline what such conditions might be.

4.3 The (Possible) Conditions

The logical place to start for making the processing operation lawful is to demand that the data be accessed lawfully. Of course, by providing a novel legal basis for the collection we are already tackling half of the problem. However, similarly to the text and data mining exception, the lawful access demand could be developed in more detail. For instance, we could determine when training data scraped from the internet or taken from websites such as Open-ML-Data containing thousands of "in-the-wild"

er-grabbing-eu-users-data-for-training-grok/ (accessed on the 20 of September 2024). LinkedIn has recently engaged in a similar practice, however, not for users located in the EU, EEA and Switzerland. See more here: Wiggers, K., 'LinkedIn scraped user data for training before updating its terms of service', TechCrunch, September 18, 2024, https://techcrunch.com/2024/09/18/linkedin-scraped-user-data-for-training-before-updating-its-terms-of-service/ (accessed on the 20 of September 2024).

[15] See footnote n(14).

datasets can be considered lawful and impose due diligence obligations onto developers using such data. Furthermore, by demanding lawful access to data used for training we would also demand that any restrictions such as social media profiles with private settings, any text-files preventing scraping and even paywalls would prohibit the data scraping regardless of the purpose of the collection.[16] Finally, this would mean that all clearly illegal sets, such as was the case with Books1, Books2 or The Pile in terms of copyright and the TDM exception, could never be used for the purposes of ML training.

Secondly, as already hinted multiple times throughout the chapter, it would be absolutely essential to only allow this exception for ML models and developers who do not aim to store the personal data or use it in a way that has as its purpose or even simply allows later reidentification. This condition would from the very start disqualify biometric identification, biometric verification and even in most cases systems such as those for making credit assessments or making price adjustments for insurance policies. On the other hand, what it would allow would be natural language processing, object recognition and semantic segmentation, disease recognition or prediction based on objective bodily factors and data, etc. regardless of whether the developers of such technologies are nonprofit or commercial entities.

Thirdly, the systems so developed could never be used to the detriment of the data subjects. This would mean that again systems such as credit scoring, but also emotion recognition, would be disqualified from relying on this exception so long as their decisions or predictions can be used in a way that can negatively impact individuals. For other types of systems, which again do not allow or enable identification after processing, this condition should be fairly straightforward. If, for example, autonomous vehicles do not store data allowing later identification of recorded persons then such recording can also not be used to identify persons being suspected of traffic violations. For other types of systems, such as LLMs and diffusion models, the situation is admittedly more complex with their inherent potential for generation of convincing yet completely false information about persons, as well as deepfakes and fake news. However, here especially we can put the burden of proof on the developers who would have to demonstrate that they undertook all reasonable and appropriate measures to prevent such uses. When malevolent actors still manage to jailbreak the models into sharing personal data, then this is the fault of that user and not the model provider or developer. This still does not mean to imply that the developer would not have to be responsible for any incurring damages.

Finally, this ties us to the other conditions which are rather procedural than substantial in nature and some of which we already mentioned. These are: reversed burden of proof, implementation of adequate and appropriate technical and organizational measures, not just when collecting the data but throughout the AI lifecycle, as well as mandating the developers engaging in such data processing to have adequate securities or insurance funds to cover damages resulting from the processing. Alternatively, one can also imagine imposing a tax on data processing or

[16] See also CNIL's *how to sheet* on scraping and relying on legitimate interests: CNIL, The legal basis of legitimate interests: Focus sheet on measures to implement in case of data collection by web scraping, 2 of July 2024, https://www.cnil.fr/en/legal-basis-legitimate-interests-focus-sheet-measures-implement-case-data-collection-web-scraping (accessed on the 19 of September 2024).

limiting the amount of profits which could be generated by relying on data altruism for data collection, both of which are wholly different rabbit holes to go down, yet they would be alternatives for compensating for the data altruism these developers benefit from when building their models. This list of conditions is far from exclusive. As mentioned for instance in Section 3.2., these conditions could be further strengthened and modified to also include putting certain data completely off limits and preventing its processing even under any such data altruism exception. Or, similarly to the text and data mining exception, by imposing different requirements on nonprofit and commercial entities. Furthermore, any of the proposed conditions would have to be further elaborated and solidified to honor the principle of transparency and legal certainty highly valued in the European Union. Nonetheless, ignoring the problem or claiming data is no longer personal because it is stored in *token* form[17] neither honors the European legal tradition and principles nor does it provide a clear and stable environment for actors relying on the current tolerance for their practices.[18]

5 Conclusion

To conclude, all new technologies make us confront certain questions that we thought we had already figured out. And while it is important to enable the development of these technologies and not prevent the progress of science, it is also equally important to address these issues before they evolve into seven headed beasts. That is before we have a whole bunch of unlawful AI systems on the market that millions of people use daily. This is why it is crucial to correctly interpret existing laws, such as the GDPR and the DGA, as well as to identify and close the persisting loopholes. This paper is an attempt to do exactly that.

The questions we considered include the limits of what users can understand and therefore lawfully consent to when it comes to complex technologies. As well as if there is a certain residuum of privacy that cannot be waived. However, even when talking about the alienable privacy, the problem of monetizing user data to achieve astronomical profits, while collecting the data through shady, dark-patterned and predominantly unlawful practices, cannot be overlooked. Adequate safeguards for using data obtained from users as a resource and the core of a business model have to be set up. On the other hand, lack of high-quality data for training many AI systems and the non-existing data economy in the EU remain important problems. Especially

[17] Again, see the recently published Discussion Paper: Large Language Models and Personal Data of the Hamburg Commissioner for Data protection and freedom of information in which they claim that LLMs do not store any personal data and therefore the data subjects rights are not relevant. See full paper here: https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Discussion_Paper_Hamburg_DPA_KI_Models.pdf.

[18] To further emphasize the urgency of creating such a stable framework see also the Open Letter ‘Europe needs regulatory certainty on AI”, published and open for signature in September 2024, https://www.linkedin.com/posts/yann-lecun_europe-needs-regulatory-certainty-on-ai-activity-7242573044739641344-p6b-/ (accessed on the 20 of September 2024).

when considering AI systems that have very little or nothing to do with individuals whose data is contained in the training sets. While this cannot be used as an excuse to ignore existing regulations, it can be used as an incentive for a meaningful conversation on how to change things. After all, the issues and problems discussed in this paper are not that novel or revolutionary. Therefore, what this paper aims to achieve is providing an overview of the current legal situation regarding transparency and AI. As well as to helping steer the discussion in a new and constructive direction.

In conclusion, this paper explained why the current legal situation threatens to suffocate AI research in Europe or at best will continue to hang over the heads of those trying to develop high-quality AI systems and ML models. The purpose of this was to demonstrate why a novel solution is necessary to remedy some of the identified issues. The framework that was proposed is not an attempt to solve all the problems, but to help instigate discussion at both the societal and political level. Because none of the problems presented here can ever be solved unilaterally. Mutual understanding, a common frame of reference and strong democratic processes are all necessary prerequisites for opening up a meaningful dialogue on how to bring the tech industry into Europe or at least not scare away what is left of it.

References

1. Appius Claudius Caecus, Roman Statesman', Britannica, https://www.britannica.com/biography/Appius-Claudius-Caecus (accessed on the 5 of January 2024)
2. McClintock, A., 'Appius Claudius Caecus and Roman law', The Encyclopedia of Ancient History, 2019, Wiley Online Library, https://onlinelibrary.wiley.com/doi/10.1002/9781444338386.wbeah30650 (accessed on the 5 of January 2024).
3. Raaflaub, K. A., 'Ancient Greece: The Historical Needle's Eye of Modern Politics and Political Thought', *The Classical World*, Vol. 109, No. 1, 2015, pp. 3-37, https://www.jstor.org/stable/24699909 (accessed on the 5 of January 2024).
4. Esperança, B., 'Analysis: The Transparency Principle in the Interpretation of the UCTD', Combat Abusive Lending, https://www.abusivelending.org/content/analysis-transparency-principle-interpretation-uctd (accessed on the 5 of January 2024).
5. Gonzales Fuster, G.,Transparency as translation in data protection', BEING PROFILED:COGITAS ERGO SUM: COGITAS ERGO SUM: 10 Years of Profiling the European Citizen, edited by Bayamlioglu, E., Baraliuc, I., Wilhelmina Janssens, L. A., and Hildebrandt, A. Amsterdam: Amsterdam University Press, 2018, pp. 52-55. https://doi.org/10.1515/9789048550180-010 (accessed on the 10 of September 2023)
6. Dorfleitner, G., Hornuf, L., and Kreppmeier, J., Promise not fulfilled: FinTech, data privacy, and the GDPR, Electron Markets 33, 33, 2023, https://link.springer.com/article/10.1007/s12525-023-00622-x , (accessed on the 24 of November 2023).
7. Judgement of the 4th of July 2023, Meta Platforms and Others v Bundeskartellamt C-252/21, ECLI:EU:C:2023:537
8. Privacy Policy, OpenAI, https://openai.com/policies/privacy-policy.

9. Duarte F., Number of ChatGPT Users (Nov 2023), EXPLODING TOPICS, https://explodingtopics.com/blog/chatgpt-users (accessed on the 24 of November 2023).
10. Introducing ChatGPT Plus, OpenAI, February 2023, https://openai.com/blog/chatgpt-plus (accessed on the 24 of November 2023)
11. Trafton, A., Legalese Fatigue: Even Lawyers Prefer Plain English, SciTechDaily, 21 June 2023, https://scitechdaily.com/legalese-fatigue-even-lawyers-prefer-plain-english/ (accessed on the 5 of January 2024)
12. Kitkowska A. et al, Enhancing Privacy through the Visual Design of Privacy Notices: Exploring the Interplay of Curiosity, Control and Affect, June 2020, Conference: USENIX Symposium on Usable Privacy and Security (SOUPS) 2020 https://www.researchgate.net/publication/342495772_Enhancing_Privacy_through_the_Visual_Design_of_Privacy_Notices_Exploring_the_Interplay_of_Curiosity_Control_and_Affect (accessed on the 24 of November 2023)
13. Privacy Policy, easyJet, YouTube, https://www.youtube.com/watch?v=o199qdIdOso&ab_channel=easyJet (accessed on the 24 of November 2023)
14. Directive 93/13 on unfair terms in consumer contracts, L 95/29
15. Regulation 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) L 119/1
16. Judgement of the 16 July 1998, Gut Springenheide, C-210/96, ECLI:EU:C:1998:369; Judgement of the 3 March 2020, Gómez del Moral Guasch, C-125/18, ECLI:EU:C:2020:138; Judgement of the 16 July 2020, Caixabank and Banco Bilbao Vizcaya Argentaria, joined cases C-224/19 and C-259/ 19, ECLI:EU:C:2020:578.
17. Judgement of the 20 September 2018, OTP Bank/Ilyés and Kiss, C-51/17, ECLI:EU:C:2018:750
18. B.M. Loos, M., 'Crystal Clear? The Transparency Requirement in Unfair Terms Legislation', *European Review of Contract Law*, vol. 19, no. 4, 2023, p. 281. https://doi.org/10.1515/ercl-2023-2018 (accessed on the 5 of January 2024).
19. Durovic, M., and Poon, J., 'Consumer Vulnerability, Digital Fairness, and the European Rules on Unfair Contract Terms: What Can Be Learnt from the Case Law Against TikTok and Meta?', *Journal of Consumer Policy* 46, 2023, p. 419, https://doi.org/10.1007/s10603-023-09546-7 (accessed on the 5 of January 2024).
20. 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, Brussels, 21.4.2021 COM(2021) 206 final and DRAFT Compromise Amendments on the Draft Report Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD)), 16.5.2023
21. Davis, F. D., Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology, MIS Quarterly, Vol. 13, №3 (1989), pp. 319–340 https://www.jstor.org/stable/249008?typeAccessWorkflow=login (accessed on the 5 of January 2024)
22. Lomas, N., 'Covert data-scraping on watch as EU DPA lays down 'radical' GDPR red-line', TechCrunch, 30 March 2019, https://techcrunch.com/2019/03/30/covert-data-scraping-on-watch-as-eu-dpa-lays-down-radical-gdpr-red-line/ (accessed on the 5 of January 2023)
23. Prezes Urzędu Ochrony Danych Osobowych, Decyzja ZSPR.421.3.2018, https://uodo.gov.pl/decyzje/ZSPR.421.3.2018(acessed on the 8 of January 2024)

24. ‘Altruism’, Cambridge Dictionary, https://dictionary.cambridge.org/dictionary/english/altruism (accessed on the 25 of December 2023)
25. Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, 4 November 1950, ETS 5
26. European Court of Human Rights, Factsheet – Personal data protection, https://www.echr.coe.int/documents/d/echr/fs_data_eng (accessed on the 24 of November 2023)
27. Charter of Fundamental Rights of the European Union 2012/C 326/02, C 326/391
28. Bock, K. and Engeler, M., Die verfassungsrechtliche Wesensgehaltsgarantie als absolute Schranke im Datenschutzrecht, *Deutsches Verwaltungsblatt*, vol. 131, no. 10, 2016, pp. 593-599. https://doi.org/10.1515/dvbl-2016-1003 (accessed on the 24 of November 2023)
29. Janeček, V., and Malgieri, G., ‘Data Extra Commercium’, 2019, in *S. Lohsse, R. Schulze and D. Staudenmayer (eds), Data as Counter-Performance—Contract Law 2.0?*, Hart Publishing/Nomos 2020, pp. 93-122., https://ssrn.com/abstract=3400620 (accessed on the 5 of January 2023)
30. Hense, P., ‘Informationsaskese vs. Meta – Perspektiven des EuGH auf datengetriebene Praktiken in der modernen Gesellschaft‘, Kommunikation&Recht 9/2023, p.556 https://online.ruw.de/suche/kur/Informationsa-vs.-Meta--Perspek-des-EuGH-auf-daten-d9da67a2a0de0196592bda003853b3b8 (accessed on the 5 of January 2023)
31. DIRECTIVE (EU) 2023/2225 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 October 2023 on credit agreements for consumers and repealing Directive 2008/48/EC, 30.10.2023, 2023/2225
32. Glowacki, M., and Stransky, S., ‘CCPA draft regulations: Privacy notices and accessibility in the employment context’, 6 July 2020, IAPP, https://iapp.org/news/a/ccpa-draft-regulations-privacy-notices-and-accessibility-in-the-employment-context/ (accessed on the 5 of January 2023)
33. Hacker, P., Engel, A., and Mauer, M., Regulating ChatGPT and other Large Generative AI Models Working Paper, version May 12, 2023, *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency 2023,* https://arxiv.org/ftp/arxiv/papers/2302/2302.02337.pdf (accessed on the 26 of November 2023)
34. Synthetic Data: The Complete Guide, datagen, https://datagen.tech/guides/synthetic-data/synthetic-data/# (accessed on the 26 of November 2023).
35. Synthetic data, TechTarget, https://www.techtarget.com/searchcio/definition/synthetic-data (accessed on the 26 of November 2023)
36. Gunasekar, S., et al., Textbooks Are All You Need, arXiv preprint arXiv:2306.11644 (2023), https://arxiv.org/pdf/2306.11644.pdf (accessed on the 26 of November 2023)
37. Hao, S., et al., "Synthetic Data in AI: Challenges, Applications, and Ethical Implications." arXiv preprint arXiv:2401.01629 (2024). https://arxiv.org/abs/2401.01629 (accessed on the 7 of January 2024)
38. Shumailov, I., et al., The Curse of Recursion: Training on Generated Data Makes Models Forget, arXiv preprint arXiv:2305.17493 (2023), https://arxiv.org/pdf/2305.17493.pdf?trk=public_post_comment-text (accessed on the 26 of November 2023)
39. Emerging Privacy Enhancing Technologies, Current Regulatory and Policy Approaches, OECD Digital Economy Papers, No. 351, March 2023, https://www.oecd-ilibrary.org/docserver/bf121be4-en.pdf?expires=1701006507&id=id&ac

cname=guest&checksum=1287047AF12DB4184F70213FCB258EA5 (accessed on the 26 of November 2023)

40. Ali, M., Synthetic data is the future of Artificial Intelligence, 18 of January 2023, Medium, https://moez-62905.medium.com/synthetic-data-is-the-future-of-artificial-intelligence-6fcfd2ce1a14 (accessed on the 26 of November 2023)
41. Veale, M, Binns, R. and Edwards, L., 'Algorithms that remember: model inversion attacks and data protection law' Phil. Trans. R. Soc.)2018) A. 376:20180083. http://doi.org/10.1098/rsta.2018.0083 (accessed on the 19 of September 2024).
42. Bitkom. Bitkom's Principles for the Data Governance Act. Position Paper. 27. Januar. Berlin: Bitkom. (2021) https://www.bitkom.org/sites/default/files/2021-01/20210127_bitkom-dga-principles-1.pdf (accessed on the 8 of September 2024).
43. Binctin, N., TDM: A CHALLENGE FOR ARTIFICIAL INTELLIGENCE, http://rida.ideesculture.fr/sites/default/files/2020-02/262-D1VA.pdf (accessed on the 19 of September 2024).
44. CEDPO AI Working Group, Generative AI: The Data Protection Implications, 16 October 2023.
45. Choi, H., et al. "The Role of Privacy Fatigue in Online Privacy Behavior." *Computers in Human Behavior*, vol. 81, 2018, pp. 42-51, https://doi.org/10.1016/j.chb.2017.12.001.
46. Ranisch, R. 'Consultation with Doctor Twitter: Consent Fatigue, and the Role of Developers in Digital Medical Ethics', *The American Journal of Bioethics*, 21(7), 2021, pp. 24–25. doi: 10.1080/15265161.2021.1926595.
47. Rasmusen, S. C., Penz, M., Widauer, S., Nako, P., Kurteva, A., Roa-Valverde, A., & Fensel, A., Raising Consent Awareness With Gamification and Knowledge Graphs: An Automotive Use Case. *International Journal on Semantic Web and Information Systems (IJSWIS), 18*(1), 2022, pp. 1-21. http://doi.org/10.4018/IJSWIS.300820
48. Commission Staff Working Document, Impact Assessment Report accompanying the DGA proposal, SWD(2020) 295 final.
49. Vardanyan, L. and Kocharyan, H. The GDPR and the DGA Proposal: are They in Controversial Relationship?. European Studies, 2022, Sciendo, vol. 9 no. 1, pp. 91-109. https://doi.org/10.2478/eustu-2022-0004
50. Baloup, J. et al. White Paper on the Data Governance Act, CiTiP Working Paper 2021, (June 23, 2021) http://dx.doi.org/10.2139/ssrn.3872703.
51. Veil, W., Data altruism: how the EU is screwing up a good idea, Algorithm Watch, January 27, 2022, https://algorithmwatch.org/en/eu-and-data-donations/ (accessed on the 20 of September 2024).