February 6th, 2023

AI-REGULATION.COM

Re-identification attacks and data protection law

By Cédric Lauradoux, Teodora Curelariu & Alexandre Lodie

To cite this article:

C. Lauradoux, T. Curelariu, A. Lodie, Re-identification attacks and data protection law, AI-Regulation.com, February 6th, 2023.

AI-Regulation.com
CHAIR LEGAL AND REGULATORY IMPLICATIONS OF ARTIFICIAL INTELLIGENCE

Re-identification attacks and data protection law

Abstract. Today's world is marked by the progress made towards the free flow of open data. This results in new challenges for data protection mechanisms, as using public datasets can lead to serious privacy breaches. To mitigate these risks, data can be anonymised. However, with the growing efficiency of re-identification attacks on anonymised data, non-personal data can be transformed into personal data. This leads to legal uncertainty for the researcher undertaking re-identification attacks. This paper tries to analyse the status of ill-anonymised data and the consequences of re-identification attacks, with regard to the GDPR. To answer these questions, we have analysed the GDPR and some national DPA's opinions and guidelines. We have drafted recommendations on the pressing need for guidelines to provide researchers carrying out such attacks with some legal certainty.

Keywords: GDPR, personal data, anonymisation, re-identification

Machine learning and other Artificial Intelligence (AI) applications are being more and more deployed. They are powered by the sharing and processing of personal data for training and validating predictive models. Hence, AI operators are calling for the free flow of data and for their free publication without restriction. This creates new challenges for data protection regulation.

A general assumption is made regarding the very nature of data, which can be divided into two categories : personal data and non-personal data. Anonymisation is the processing of data which transforms personal data into non-personal data by removing the identifiable features of a person, which enables data controllers to publish a dataset. Many propositions have been made to anonymise a dataset. However, there is a consensus to say that perfect anonymisation does not exist: there are still some risks of re-identification.

Re-identification attacks are becoming more and more sophisticated and efficient. There are several forms of re-identification on anonymised (or pseudonymised) datasets: linkage and inference attacks. The former attacks, also called database crossing, are made possible by the availability and accessibility of online datasets containing large amounts of personal data. The latter attacks have benefited from the development of AI. It is now possible to infer sensitive data from a dataset which did not contain any initially. Therefore, the likelihood and the effectiveness of re-identification attacks should not be underestimated. Data controllers need to be ready to handle such threats and their consequences, *i.e.* data breaches and risks for data subjects privacy.

Indeed, creating a truly anonymous dataset with a certain utility in order to prevent further identification is impossible so far. Anonymisation may mitigate the risks as its goal is to achieve irreversible de-identification, but linkage attacks and the use of other available sources of information may prevent proper irreversible anonymisation. Moreover, even if the anonymised data in question might not contain personal information,, it becomes easier and easier for anonymised data to be "transformed" into personal data by using linkage methods or by de-anonymising datasets, as stated by Working Party 29[1].

Re-identification attacks involve a wide array of legal questions in particular with regard to data protection laws. The main challenge is to know what legal framework applies to the data which was anonymised in the first place and then re-identified. Let assume that a controller Alice has published an anonymised dataset. Later, another person, Eve, is able to re-identify the individuals involved in Alice's dataset. We aim at answering the following questions:

- What is the legal status of the data contained in Alice's dataset before and after Eve's attack?
- What are the responsibilities of Alice and Eve with regard to subjects' privacy?
- Does European data protection law expressly address re-identification attacks?

This article aims at assessing these issues about the status of ill-anonymised data according to data protection law and the consequences that a re-identification attack might raise for the actors concerned, especially for the researcher carrying out such re-identification.

This paper is the result of the work of a multidisciplinary team, involving legal scholars and a computer scientist. First, the technical background of anonymisation methods and some re-identification attacks is discussed. Second, some legal aspects are developed in order to apprehend what the legal consequences attached to a successful re-identification attack are. Finally, recommendations are made towards local DPAs and the EDPB as to provide legal certainty to researchers undertaking re-identification attacks.

I. What is data re-identification?

Before considering the legal consequences related to re-identification and the responsibility of the data controller(s), it is worth giving some insights about what anonymisation and re-identification really are and how they can be carried out in practice. A data controller can perform three types of processing with respect to the definition of personal data and non-personal data:

1. Processing of personal data which remains personal data.
2. Processing of personal data which are transformed into non-personal data.
3. Processing of non-personal data which are transformed into personal data.

We call the first type neutral processing while the second is anonymisation and the last one is called a re-identification attack. Neutral processing and anonymisation are both discussed in the GDPR. Re-identification and neutral processing stay in the scope of the GDPR while anonymisation lies outside the scope of the GDPR, as it will be further explained in detail[2]. The identifiability of a person after a certain processing is the key element to determine if a processing results in a successful anonymisation process.

[1] Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation (WP 203, 2013, available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf , last accessed on 25th January, 2023

[2] See below, Part II

1. Anonymisation

We first discuss the finalities of anonymisation then we focus on the different techniques of anonymisation. Finally, we analyse how anonymisation is assessed.

Anonymisation has two opposing goals: privacy and utility. It aims to minimise the risks of subjects re-identification in case of an attack. The utility is a measure of how useful the anonymised dataset is compared to the original dataset. There are distortions between the anonymised dataset and the original data. Computations made on the original dataset must be unchanged or close compared to those on the anonymised dataset. There are naive solutions to achieve anonymisation with perfect privacy (exclusive) or utility. Anonymisation with perfect privacy substitutes all the personal data by random ones. However, such a solution has no utility. Anonymisation with perfect utility does not modify anything from the original dataset in the anonymised dataset. Computation remains unchanged but this ``anonymised dataset'' does not protect the subjects' privacy. Therefore, the design of an anonymisation scheme requires finding an appropriate trade-off between data privacy and utility. Privacy scholarship[3] has established that it is not possible to achieve anonymisation with both perfect data protection and perfect utility. Data protection authorities have acknowledged situation[4].

From a technical perspective, anonymisation is based on three techniques: deletion, randomisation and generalisation. Randomization adds noise to the original database by substituting values by random ones for instance. Generalisation substitutes values in the original data by more general values. The readers can consult ENISA report on data protection engineering[5] or the annex of Article 29 Data Protection Working Party's opinion on anonymisation techniques[6] for more details. Several anonymisation models have been proposed in the past like anonymity-set-size, k-anonymity, l-diversity or differential privacy to understand how deletion, randomisation and generalisation need to be applied to a dataset in order to prevent re-identification. Those models have two goals: (i) they aim at providing guarantees on the uncertainty that an adversary will have to re-identify and (ii) they provide an effective process to reach these guarantees. Many models have limits and re-identification was proven possible in many cases despite the use of state-of-the-art anonymisation techniques as explained later in this section.

One has to look at all means reasonably likely to be used by anyone to identify an individual[7] in order to evaluate the performance of an anonymisation process. Article 29 Data Protection Working Party[8] recommends to consider and evaluate the following three risks: singling out, linkability and inference. These residual risks occurring when using anonymisation need to be quantified and minimised. As the re-identification techniques are still evolving, it is difficult to know if the current methods used to assess the security of an anonymisation process are sufficient or not.

[3] Ohm Paul, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." Ucla L. Rev. n°57 (2009) : 1701; Dwork, Cynthia, Adam Smith, Thomas Steinke, Jonathan Ullman, "Exposed! A Survey of Attacks on Private Data." Annual Review of Statistics and Its Application, (2017).

[4] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques.

[5] ENISA, "Data Protection Engineering : From Theory to Practice", January 2022, available at: https://www.enisa.europa.eu/publications/data-protection-engineering, last accessed on January 25th, 2023.

[6] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques.

[7] See Recital 26 of the GDPR.

[8] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques.

To conclude on this brief introduction to anonymisation, there is a clear distinction between pseudonymisation and anonymisation from the GDPR's perspective. Pseudonymisation is a neutral processing, i.e. pseudonymised data are personal data. However, there is still the problem to qualify if a data protection technique is qualified as a pseudonymisation or an anonymisation technique. There are still many discussions concerning data protection[9].

2. Historical background of data re-identification

The following table aims at giving some insight on the most consequent re-identification attacks.

Year	Dataset creator	Type of data	Defence used	Attack type	Authors
1997	NAHDO[10] GIC[11] Cambridge Massachusetts	Hospitalisation records *(ZIP code, birth date, gender)* Medical records in the GIC data *(Ethnicity, Visit date, Diagnosis, Procedure, Medication, Total Charges)* Voter registration data *(Name, Address, Date registered, Party affiliation, Date last voted)*	Pseudo.	Crossing databases	Latanya. Sweeney[12]
2000	NAHDO Cambridge Massachusetts	Hospitalisation records *(ZIP code, birth date, gender)* Voter registration data	Pseudo.	Crossing databases (census)	Latanya Sweeney[13]
2006	AOL	Users' search queries	Pseudo.		Michael Barbaro, Tom Zeller Jr[14]

[9] See below, Part II

[10] The National Association of Health Data Organizations

[11] The Group Insurance Commission

[12] Latanya. Sweeney. "k-anonymity: a model for protecting privacy". *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570

[13] Latanya Sweeney. "Simple demographics often identify people uniquely". *Health* (San Francisco),671(2000):1-34, 2000

[14] Michael Barbaro, Tom Zeller Jr, "A Face Is Exposed for AOL Searcher No. 4417749", The New York Times, August 9, 2006, available at : https://www.nytimes.com/2006/08/09/technology/09aol.html, last accessed on January 25th, 2023.

2007	Netflix	Users’ movie preferences	Pseudo.	Crossing databases (IMDB)	Narayanan, Arvind, and Vitaly Shmatikov[15]
2008	Cabspotting	Taxi trajectory (GPS coordinates)	Pseudo.	Point of interests discovery	Gambs, Sébastien, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez[16]
2017-2018	Strava	Users’ trajectories (GPS coordinates)	Privacy area	Regression	Dhondt, Karel, Victor Le Pochat, Alexios Voulimeneas, Wouter Joosen, and Stijn Volckaert[17]
2018-2019	Swiss Federal Supreme Court, Swiss Federal Administrative Court Swiss Federal Office of Public Health	Court decisions Drugs data	Pseudo.	Crossing databases + Web scraping	Kerstin Noëlle Vokinger / Urs Jakob Mühlematter[18]
2021	edX (Harvard…)	Students enrolled in edX courses	k-anonymity	Crossing databases	Aloni Cohen[19]

As the table above illustrates, re-identification attacks put personal and non-personal data at stake, despite the defence used.

[15] Narayanan, Arvind, and Vitaly Shmatikov. "How to break anonymity of the netflix prize dataset." *arXiv preprint cs/0610105* (2006).

[16] Gambs, Sébastien, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. “De-anonymization attack on geolocated data”. *Journal of Computer and System Sciences* 80, no. 8 (2014): 1597-1614.

[17] Dhondt, Karel, Victor Le Pochat, Alexios Voulimeneas, Wouter Joosen, and Stijn Volckaert. “A Run a Day Won't Keep the Hacker Away: Inference Attacks on Endpoint Privacy Zones in Fitness Tracking Social Networks” In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 801-814. 2022.

[18] Kerstin Noëlle Vokinger / Urs Jakob Mühlematter, *Re-Identifikation von Gerichtsurteilen durch «Linkage» von Datenbanken),* in : Jusletter 2 septembre 2019

[19] Aloni Cohen, “Attacks on deidentification’s defenses”. In 31st USENIX Security Symposium (USENIX Security 22), pages 1469–1486, Boston, MA, August 2022. USENIX Association

The first historical re-identification attack was undertaken by computer scientist Latanya Sweeney in 1997 by crossing databases. She managed to identify the Governor of Massachusetts by matching hospitalisation records with voter registration records, putting at stake data protection techniques used by American public administrations[20]. A couple of years later, she came to the conclusion that 87% of the US population could potentially be identified by their ZIP code combined with additional information, such as gender and birth date[21].

Later on, AOL publicly released a pseudonymised dataset of its users' Web search query logs. Information about users' search history (including queries about political views and medical conditions) was accessed[22]. One year after, Netflix also publicly released a pseudonymised dataset containing information about its users' movie ratings. Similarly to the AOL use-case, the dataset in question did not include any personally identifying information, but, when cross-correlated with auxiliary information available from other sources, re-identification of the said dataset was possible. Both examples had serious legal consequences, as class action lawsuits were filed against them on data privacy grounds[23]. Arguments by plaintiffs in the Netflix class action included, among others, the violation of privacy by "*the disclosure to third parties of sensitive and/or personal identifying information*"[24] derived from the activity of its users. The contents shared by Netflix included the subscribers' renting history and habits, communications, rating videos information, without notice to or consent by their subscribers, having led to the disclosure of sensitive information, such as sexual orientations.

All of the examples mentioned above worked by cross-referencing the pseudonymised dataset with auxiliary knowledge obtained by other means. However, crossing databases is not the only mechanism allowing re-identification, as the Cabspotting case illustrates: the dataset containing taxi trips was insufficiently protected by the pseudonymisation algorithms, leading to access to GPS coordinates using attacks based on Points of Interest.

Using pseudonymisation and removing identifying information in order to protect data might not be enough to ensure data protection. The main issue about pseudonymisation is that it is a reversible operation, meaning that the data that has been pseudonymised can still be traced down using third-party data. From a legal point of view, data protection legislations apply to pseudonymised data, as the data in question is not meant to be irreversibly altered, and thus remains personal.

On the contrary, anonymisation operations are not reversible (at least in theory), as they are aimed at making it impossible to identify individuals. The insufficiency of anonymisation was illustrated by the edX attack undertaken by Aloni Cohen, who stated that current anonymisation techniques might not be sufficient to meet the legal bar of irreversible

[20] Latanya Sweeney, "k-Anonymity: A Model for Protecting Privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, n°10 (2002): 557, 558–59

[21] Latanya Sweeney, "Simple Demographics Often Identify People Uniquely". Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000, 1-34

[22] Michael Arrington (August 6, 2006). "AOL proudly releases massive amounts of user search data". TechCrunch. Archived from the original on August 12, 2006. Retrieved August 7, 2006

[23] AOL Class Action Settlement | Landwehr v. AOL Search Data Privacy Lawsuit, https://classactionlawsuitsinthenews.com/class-action-lawsuit-settlements/aol-search-data-privacy-class-action-settlement-landwehr-v-aol/, last accessed on 25th January, 2023

[24] Jane Doe v. Netflix, Inc. et al, US District Court of the Northern District of Indiana, available at https://www.wired.com/images_blogs/threatlevel/2009/12/doe-v-netflix.pdf, last accessed on 25th January, 2023

de-identification[25]. EdX publicly released data, which was then combined with auxiliary information found on LinkedIn. Identifying information about individuals who registered for edX courses or who started but failed was found[26]. This operation emphasised that advanced anonymisation techniques (k-anonymity for instance) that transforms data to make an individual indistinguishable from others in a given dataset is what makes the method vulnerable. It is wrongly perceived that if the data does not contain direct identifiers (if it is anonymised), it cannot be tracked down to a particular individual.

II- The legal consequences of a re-identification attack

Before even considering the legal risks attached to a re-identification attack, we need to take a step backwards and analyse the legal status of an anonymised dataset published on the open web. The issue at stake is to know what is the legal status of anonymised datasets in order to understand what might be the responsibilities of each party who has processed it. Eventually, it is worth considering whether a re-identification can be considered as a further processing.

1. The legal status of public anonymised datasets and the responsibility of their publishers

The edX attack[27] which reidentified anonymised data by combining datasets has shown that anonymisation does not prevent re-identification in an absolute manner. From a legal point of view and contrary to pseudonymisation, data protection legislations no longer apply when the data in question is (considered to be) anonymised, as there is no processing of personal data. In such a case,*"personally identifiable information is irreversibly altered in such a way that the subject of the personally identifiable information can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party"*[28].

Therefore, with regard to the GDPR, anonymised data seemed to lie outside the scope of the regulation. Indeed, recital 26 provides that *"(t)he principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes"*[29]. In other words, when data are anonymised, they are no longer

[25] Rob Mitchum, "New kind of attack called 'downcoding' demonstrates flaws in anonymizing data", Techexplore, University of Chicago, October 10, 2022, available at: https://techxplore.com/news/2022-10-kind-downcoding-flaws-anonymizing.html , last accessed on 25th January, 2023

[26] Aloni Cohen."Attacks on deidentification's defenses". In 31st USENIX Security Symposium (USENIX Security 22), pages 1469–1486, Boston, MA, August 2022. USENIX Association

[27] See above, part I.

[28] ISO 29100: 2011, as cited in Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques

[29] See Recital 26 of the GDPR.

considered “personal data” and the data controller is therefore not bound by the GDPR anymore[30], but by another EU regulation[31].

It is thus critical to understand what the definition of anonymisation is because this definition bears huge legal consequences. Unfortunately, the GDPR does not specify what techniques are compliant with the anonymisation requirement. Recital 26 only underlines that “*(t)o determine whether a natural person is identifiable, account should be taken of 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. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments”*[32]. These various assertions leave great room for interpretation[33].

As a matter of fact, it is up to the data controller to ensure that he/she has properly anonymised the dataset. This anonymisation means that individuals are made unidentifiable. However, since technology keeps improving, as evidenced for instance by the development of machine learning techniques, no anonymised dataset can be considered as fully secure. If there is always a risk of re-identification, it does mean that there is no such thing as anonymised data and that GDPR must always apply. According to some scholars this view expresses an “absolute approach” according to which as long as there is a risk of re-identification, data must be considered as “personal data” and any data breach should be attributed to the data controller[34]. However, the GDPR seems to acknowledge to a certain extent a relative approach since it directly refers to the “*means (..) likely to be used to identify the natural person”*[35]. In other words, “*this approach only considers the real knowledge, the necessary means and effort that are "reasonably likely" to be used by the data controller to identify a person*”[36]. From this perspective, once the data controller has taken all the organisational measures to anonymise data, these data are not regulated under the GDPR anymore and thus the data controller cannot be held liable for any further data breach.

The logic behind these provisions is that once data are appropriately anonymised, they are no longer linkable to a specific subject and thus they are no longer considered as personal data. This logic is based on the premise that complete anonymisation is technically feasible. For instance, Working Party 29 seems to acknowledge this view as it considers that “*the result of anonymisation, as a technique applied to personal data, should be, in the current*

[30] Karine Bannelier, Anaïs Trotry, “What is ‘data’? Definitions in International Legal Instruments on Data Protection, Cross-Border Access to Data & Electronic Evidence”, ai-regulation.com, January 10th, 2023.

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

[32] See Recital 26 of the GDPR

[33] Bygrave, Lee.A, *Data Protection Law: Approaching Its Rationale, Logic and Limits*, Kluwer Law International, 2002

[34] Lukas Helminger, Christian Rechberger, “Multi-Party computation in the GDPR*”, IACR Cryptol.* ePrint Arch. 2022: 491 (2022)

[35] See Recital 26 of the GDPR

[36] Gerald Spindler, Philipp Schmechel, “Personal Data and Encryption in the European General Data Protection Regulation”, *Jipitech*, 2016, pp. 164-177.

state of technology, as permanent as erasure, i.e. it should make it impossible to process personal data"[37].

The reality is, and history has shown that, even an anonymised dataset can be de-anonymised (or re-identified). We have provided a few famous examples of successful re-identification attacks[38]. To summarise it can be considered that re-identification turns non-personal data into personal data, so it changes the legal framework applicable to the data.

The status of the data appears to be in a legal grey zone in such a scenario. As a matter of fact, articles 33 and 34 of the GDPR provide for a series of obligations to be fulfilled by the data controller in case of a data breach[39]. On the other hand, these obligations do not seem to fit our 're-identification attack' scenario because once data have been anonymised, the GDPR no longer applies and thus the data controller is not subject to these obligations anymore. So if a data controller publishes an anonymised dataset online, he/she should not - according to this logic - be held liable for any further breach of these data. The debate deserves to be kept open, because, it would seem logical that a data controller publishing data online (albeit anonymised) shares a part of the responsibility in case of a further data breach. Besides, the French 'Commission Nationale de l'Informatique et des Libertés'[40] stated that in a situation where a data breach occurs because a released dataset actually contains personal data, "*the dataset in question should therefore be removed as soon as possible*"[41]. This would mean that data controllers are still responsible for the data they published even when the GDPR no longer applies. This stance should be further detailed by the CNIL to provide researchers and data controllers with legal certainty[42].

2. **The legal status of the researcher undertaking reidentification attacks**

The debate then shifts as to whether the person who succeeds in re-identifying data must be considered as a (new) data controller. In view of the above it would seem logical to conclude that the status of the data changes again once a re-identification attack succeeds because data subjects become identified or at least identifiable. It is worth recalling that reidentification is a process enabling the transformation of a non-personal data into personal data[43]. As stated by the Norwegian Data protection authority, "*(i)f someone should succeed in re-identifying the data, and this results in personal data being processed, the organisation responsible for the data must assume the role of data controller for them, in accordance with the Personal Data Act*"[44]. This conclusion seems to be in line with article 4 (7) of the GDPR[45]

[37] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, last accessed on January 25th, 2023,

[38] See above, part I.

[39] See articles 33 and 34 of the GDPR.

[40] The CNIL is the French Data Protection Authority, in charge of investigating the violations of the GDPR.

[41] CNIL, « L'anonymisation des données, un traitement clé pour l'open data », October 17th, 2019, available at : https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data , last accessed on 25th January, 2023

[42] See below, part III.

[43] See above, Part I.1.

[44] Datatilsynet, "The anonymisation of personal data", available at: https://www.datatilsynet.no/en/regulations-and-tools/reports-on-specific-subjects/anonymisation/?print=true , last accessed on 25th January, 2023

[45] See article 4(7) of the GDPR.

since a researcher re-identifying data determines the purposes and means of the processing of personal data. Indeed, he/she uses re-identification tools to reveal anonymisation vulnerabilities which may lead to massive data breaches.

Furthermore, according to the GDPR, processing of data involves (among other actions) collecting, consulting or using data[46]. From this perspective a person who re-identifies individuals from an anonymised dataset should be considered as a data processor besides being a data controller. The purpose of this data processing lies in the scientific progress that computer scientists achieve by discovering new weaknesses of an anonymisation process used to create the online dataset. However, a data controller should process personal data in a lawful manner so before processing any personal data, a researcher willing to launch a re-identification attack must ensure that he/she can benefit from a legal basis to do so.

3. The legal basis to process data when re-identifying 'anonymised' datasets for scientific purposes

Since a computational researcher re-identifying data should logically be considered as data controller[47], he or she must act by complying with at least one of the legal bases provided for in the GDPR. Indeed, the first principle relating to processing of personal data as laid down in Article 5 of the GDPR is that "*data shall be (...) processed lawfully.*"[48].

The purpose of this processing of personal data can be described as a scientific purpose since the main aim is to reveal data security vulnerabilities and thus to protect data subjects' personal data and privacy.

From this background, the French DPA released guidelines on the legal regime applicable to data processed for scientific purposes. The guidelines identify as possible legal bases, the consent of subjects, the performance of a task carried out in the public interest or the legitimate interests pursued by the controller[49]. However, we will see that each of these candidates may involve interpretation issues or do not fit the reality of re-identification attacks carried out for scientific purposes.

First, consent cannot be the legal basis for processing data when launching a re-identification attack. As a matter of fact, a researcher carrying out such an attack is unaware of who the data subjects are, since the main aim of his/her operation is to try to identify data subjects from an 'anonymised' dataset. Such a legal basis is thus inoperative to provide a clear framework.

The CNIL's guidelines also mention the performance of a task carried out in the public interest. However, once again, it is unclear whether such a legal basis is fit for purpose in such a scenario since the GDPR requires that "*(w)here processing is carried out in accordance with a legal obligation to which the controller is subject or where processing is*

[46] See article 4 of the GDPR.

[47] Cf supra, I.B.

[48] See article 5 of the GDPR.

[49] CNIL, « Régime juridique applicable aux traitements poursuivant une finalité de recherche scientifique (hors santé) », pp. 2-3, available at : https://www.cnil.fr/sites/default/files/atoms/files/consultation_publique_-_presentation_du_regime_juridique_applicable_aux_traitements_a_des_fins_de_recherche.pdf , last accessed on 25th January, 2023

necessary for the performance of a task carried out in the public interest or in the exercise of official authority, the processing should have a basis in Union or Member State law"[50].

In other words, member states' law should expressly provide for the data processing carried out by the researcher willing to re-identify data. The question goes as to whether this law must be specific or whether a general statute of the researcher under domestic law could be sufficient as well as a mere transposition of the GDPR into domestic law[51].

For instance article L 112-1 of the French research code lists in a broad manner the objectives of public research fulfilled by scholars when carrying out their research work[52].

It is undeniable that a computer scientist performing a re-identification attack contributes to the progress of research in the field of cybersecurity and data protection. As such, a researcher and member of a public-funded institution under the authority of the Ministry of Higher Education and Research can be considered as exercising a task carried out in the public interest. The question remains to determine to what extent domestic law provisions must be specific to authorise particular data processing such as re-identification attacks for scientific purposes.

On this specific topic, the GDPR tends to suggest that the legal basis authorising a data processing in the public interest must be specific since it "*may contain specific provisions to adapt the application of rules of this Regulation, inter alia: the general conditions governing the lawfulness of processing by the controller; the types of data which are subject to the processing; the data subjects concerned; the entities to, and the purposes for which, the personal data may be disclosed; the purpose limitation; storage periods; and processing operations and processing procedures, including measures to ensure lawful and fair processing such as those for other specific processing situations as provided for in Chapter IX. (...)*"[53]. As such a general statute on public research does not seem to satisfy the requirements laid down by article 6§3 of the GDPR to authorise data processing.

The last legal basis likely to authorise a researcher to undertake a re-identification attack is the legitimate interest of the data controller. However, this legal basis seems to be inoperative for our case study since "*this basis applies only to private entities*"[54]. Indeed, recital 47 of the GDPR provides that "*(g)iven that it is for the legislator to provide by law for the legal basis for public authorities to process personal data, that legal basis should not apply to the processing by public authorities in the performance of their tasks*"[55].

Interestingly, some universities have published their own guidelines to consider under what grounds their agents could process data. For instance, the University College London claims on its official website that "*(a)s a public authority, most of UCL's processing will be undertaken using Article 6(1)(e) above, the 'public task' condition. This applies when the*

[50] See recital 45 of the GDPR.

[51] See for instance article 78 of the French Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés and article 100-1 of the Décret n° 2018-687 du 1er août 2018 pris pour l'application de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, modifiée par la loi n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles

[52] See article L 112-1 of the French research code, available at: https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000027747800/, last accessed on 25th January, 2023

[53] See Article 6§3 of the GDPR.

[54] Gabe Maldoff, "How GDPR changes the rules for research", *IAPP*, April 19th, 2016, available at: https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research/, last accessed on 25th January, 2023

[55] See Recital 47 of the GDPR.

processing is necessary for UCL to perform a task in the public interest. Examples include most of UCL's research, teaching and learning activities – we can clearly demonstrate a 'public task' basis for these because performing such tasks is a core part of UCL's Charter and Statutes"[56].

The 'public interest' legal basis seems therefore to be fit for purpose when considering the re-identification of anonymised datasets carried out by a researcher in the field of cybersecurity in the fulfilment of his or her tasks.

If researchers carrying out re-identification must be considered as data controllers and can process data lawfully, the question turns to what their obligations vis-à-vis data subjects are. If the application of the GDPR raises many questions with regard to such a scenario, it would be impossible to deal with every single one of them. We have selected a few issues.

In particular, in the scenario that we are considering, it seems very difficult for the data controller (the researcher) to comply with the right of data subjects to information as provided for by article 14 of the GDPR[57]. Indeed, when re-identifying an anonymised dataset, researchers do not know who the data subjects are, so they cannot inform them about the data processing carried out. They can only inform them *a posteriori*, which is not the right way to proceed since "*(n)otice should be provided at the time when the data is first collected, and it must include the controller's identity and contact information*"[58].

However, Article 14 paragraph 5 provides for some exemptions, in particular data controllers do not have to provide a notice when the situation makes it impossible or too complex. The same goes when such a requirement is likely "*to render the processing impossible or seriously impair the achievement of the objectives of that processing*" [59].

By virtue of this article a computer scientist undertaking re-identification attacks would not be constrained to inform people since it would be impossible because the data controller does not know the exact nature of the data processed, nor who the data subjects actually are.

[56] "Practical Data Protection Guidance Notice : Legitimate interests as a lawful basis for processing personal data", University College London, available at: https://www.ucl.ac.uk/data-protection/guidance-staff-students-and-researchers/practical-data-protection-guidance-notices/legitimate , last accessed on 25th January, 2023

[57] Article 14 of the GDPR reads as follows:

"1. Where personal data have not been obtained from the data subject, the controller shall provide the data subject with the following information:

(a) the identity and the contact details of the controller and, where applicable, of the controller's representative;
(b) the contact details of the data protection officer, where applicable;
(c) the purposes of the processing for which the personal data are intended as well as the legal basis for the processing;
(d) the categories of personal data concerned;
(e) the recipients or categories of recipients of the personal data, if any;
(f) where applicable, that the controller intends to transfer personal data to a recipient in a third country or international organisation and the existence or absence of an adequacy decision by the Commission, or in the case of transfers referred to in Article 46 or 47, or the second subparagraph of Article 49(1), reference to the appropriate or suitable safeguards and the means to obtain a copy of them or where they have been made available.

[58] Gabe Maldoff, "How GDPR changes the rules for research"

[59] See article 14 §5 (b) of the GDPR.

Furthermore, under the GDPR, data subjects have a right to object to their data processing, they are also granted a right of access, a right to rectification and a right to restriction of processing. However, when a data processing for scientific purpose is involved, the data controller may be exempted from compliance with these obligations[60]. It means that a researcher undertaking a re-identification attack would not have to protect all these rights. However, these exemptions must be provided by European or domestic law, besides they are not absolute and "*must be necessary for the fulfilment of [the research] purposes*"[61].

In a similar vein, article 89 of the GDPR also provides that, in order to comply with the data minimisation principle, data controllers may have to use pseudonymisation[62]. It would be a non-sense to ask a researcher carrying out a re-identification attack for a scientific purpose to use pseudonymisation whereas the specific aim is to de-anonymise data to reveal vulnerabilities in anonymisation techniques.

In brief, although the 'mission of public interest' can suit re-identification attacks carried out for scientific purposes as a legal basis, there are still some legal grey zones to clarify. In particular, it is unclear whether domestic or EU law must provide for specific data processing or whether broad provisions on research statutes in domestic law are sufficient to authorise such re-identification attacks. Besides, the extent to which computer scientists can derogate from data subjects' rights to object, right of rectification, or erasure when undertaking re-identification attacks remains to be set.

4. Re-identification attacks considered as a "further processing"

From a broader perspective, there is a debate on the exact nature of anonymised data. This question is critical when considering accountability issues. If an anonymised dataset is not subject to the GDPR anymore it means that the first data controller (the person who anonymised the dataset and who published it) is no longer responsible under the GDPR. From this perspective the re-identifier must be considered as a new data controller.

On the other hand of the spectrum, one might consider that if a dataset has been re-identified it would mean that the initial data controller has not complied with its obligations under Recital 26 of the GDPR and that it must therefore be held responsible for any further data breach. In such a scenario the full responsibility of any violation of the GDPR must be borne on the first data controller (the 'anonymiser').

Another option, which lies in the middle would tend to consider the action of re-identifying data as a further processing. From this background, article 5. 1 (b) of the GDPR provides that "*further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation')*"[63].

However, it is unclear whether re-identification attacks can be considered as a further processing. More specifically it is not clear whether a data processing can be considered as a further processing when there are two different data controllers. One may consider that it is

[60] See article 89 of the GDPR.

[61] See Gabe Maldoff, "How GDPR changes the rules for research"

[62] See article 89 of the GDPR.

[63] See Article 5 of the GDPR.

the case since the French ‘loi informatique et libertés’ provides that data subjects “*have the right to object, free of charge, to their data being used for canvassing purposes, in particular commercial canvassing, by the current data controller or that of a further processing*”[64]. In that respect, an Opinion by the Working Party 29 claims that “*it will also be relevant to distinguish between situations where the further processing will be carried out by the initial data controller and those where personal data will be transferred to a third party*”[65]. In any case, if re-identification attacks were to be considered as a further processing data controllers would still have to comply with requirements laid down by the GDPR such as the existence of a legal basis[66]. Such a view would deserve to be better explained by data protection authorities to avoid any legal uncertainty.

It is for all these reasons that we call upon data protection authorities to publish guidelines on this specific issue, to provide computer scientist researchers with legal certainty in the pursuance of their mission. From this perspective we have drafted a series of recommendations to data controllers and data protection authorities to address these uncertainties.

III- Our recommendations

Re-identification attacks undertaken by researchers for scientific purposes thus raise a wide array of legal issues, such as the status of the anonymiser, the status of the re-identifier, the status of the data itself (Is it subject to the GDPR? Does it cease to be? When?) that needs clarification. Utmost care should be taken when a researcher is willing to re-identify data since there might be some penal risks doing so. It is for all these reasons that we address these following recommendations:

1) Recommendations to data protection authorities and the European Data Protection Board :

- A need for clarification of the legal obligations of the re-identifier with regard to the GDPR

As mentioned previously, there is a need that the EDPB and/or local data protection authorities take a position on what legal basis can be used to process anonymised data to re-identify them, for scientific purposes. These actions contribute to reveal vulnerabilities online and to create a safe digital environment by identifying ill-anonymised datasets so they should legitimately be considered as fulfilling the ‘task carried out in the public interest’ legal basis. In any case, researchers cannot be left in the dark and act without clear legal guidance. The issuance of such guidelines could also clarify more broadly the exercise of the right of data subjects’ rights mentioned in articles 15, 16, 18, 21 of the GDPR in such a scenario.

[64] See article 38 of the Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés

[65] Article 29 Data Protection Working Party, “Opinion 03/2013 on purpose limitation”, April 2nd, 2013, p. 29, available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf, last accessed on 25th January, 2023

[66] *Ibid*, p. 28.

- A need for clarification of the first and second data controllers' responsibility

The publication of guidelines could provide data controllers with guidance on how they must implement organisational measures and cooperate with data protection authorities when identifying vulnerabilities. In other words, once a dataset has been de-anonymised, what are the obligations of the (new) data controller? Under normal circumstances a data controller which notes a data breach has a duty to report it to its data protection authority and must then inform data subjects that their data have potentially been leaked.

In our scenario the data breach is provoked by the researcher, so what steps should he/she take to ensure that his/her findings will not be harmful for data subjects' privacy? On this topic, Article 33 of the GDPR provides that there is a need for the data controller to notify the data breach to the competent authority not later than 72 hours[67]. This article seems to apply when the data controller is a victim of the data breach or when he contributed, by his negligence, to leak data. In any case, clear guidelines could be beneficial to ensure that the researcher re-identifying data adopts the right behaviour once he/she has re-identified data.
As regards the obligations of the data controller to report a data breach, the French DPA stated in its guidelines that when a published dataset online is not well anonymised or has been subject to a successful re-identification, it becomes "*necessary to remove the dataset in question as soon as possible and to file a data breach notification with the competent data protection authority if the breach is likely to result in a risk to the rights and freedoms of the data subjects*"[68]. In our opinion there is a *contradictio in terminis* between this statement and recital 26 of the GDPR. Indeed, if the first data controller (the anonymiser) is not bound by the GDPR anymore because he/she has anonymised data, it would be illogical to hold him/her accountable for any further successful re-identification action. It would mean that the application of the GDPR is temporarily suspended rather than terminated.

Interestingly, the CNIL seems to feed this legal uncertainty since it claims in other guidelines that the GDPR " *(...) no longer applies at the end of the process, as the dissemination or re-use of anonymised data no longer has any impact on the privacy of the data subjects*"[69]. If the anonymisation process has a beginning and an end it means that when the anonymiser has anonymised and published data he/she has discharged his/her responsibility.

There is thus a very pressing need for clarity as to who is in charge of reporting a data breach and who is responsible for a successful re-identification process, which can be considered as a data breach. Is it the person who anonymised the data and who failed to make the anonymisation fully infallible (the GDPR does not seem to adopt an absolute approach) or is it the person who successfully re-identified data (based on the principle that if he/she had not undertaken this re-identification process, data would have remained anonymised)?

[67] "(i)n the case of a personal data breach, the controller shall without undue delay and, where feasible, not later than 72 hours after having become aware of it, notify the personal data breach to the supervisory authority competent in accordance with Article 55, unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons. Where the notification to the supervisory authority is not made within 72 hours, it shall be accompanied by reasons for the delay"

[68] CNIL, « Recherche scientifique (hors santé) : enjeux et avantages de l'anonymisation et de la pseudonymisation », January, 31st, 2022, available at : https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation , last accessed on 25th January, 2023

[69] *Ibidem*.

- The need to take into account criminal risks related to a violation of data protection laws

This recommendation is related to the previous one and questions once again the responsibility for re-identification actions.
If the researcher processes data in an unlawful way (in absence of a clear legal basis), it can be considered not only as a violation of the GDPR but also a violation of Member States' criminal law[70]. It means that in theory, a researcher willing to re-identify data could be held responsible for a felony under domestic law because it has processed personal data without a proper legal basis or without informing data subjects. It would be very paradoxical since, by doing so, researchers are willing to protect data subjects' privacy. Hence, the pressing need to clarify the responsibility of researchers as regards the GDPR. Besides, it is worth mentioning that States are adopting legislations to protect 'white hats'[71], i.e ethical hackers[72] against legal pursuit when they reveal vulnerabilities. It remains to be seen whether researchers carrying out re-identification attacks on a publicly available dataset can benefit from this kind of protection.

2) Recommendations to researchers willing to re-identify data

- The need to act with a clear legal basis

As previously mentioned, it is very likely that re-identification attacks can be considered as a new data processing and that researchers may be considered as data controllers. In that respect they must act in compliance with the GDPR and identify a legal basis to process data. Since consent cannot be invoked (because the re-identifier does not know who data belongs to) and the 'legitimate interest' exception applies only to private entities, a researcher from a public institution should invoke the 'task carried out in the public interest' exception, keeping in mind that such an exception should also be based on a domestic or European law provision. In any case, the researcher should clearly mention the legal basis that he/she relies on, as for instance the UCL did[73]. It is worth recalling that the (un-)lawfulness of data processing in a certain region of the world does not presume anything about the lawfulness of such a processing in other States. Indeed, since data processing might involve various legislations (due to the nationality of data subjects), a data controller risks legal proceedings abroad, as the example of the firm Cambridge Analytica demonstrated it[74].

[70] For instance article 226-18 of the French criminal code provides that "(c)ollecting personal data by fraudulent, unfair or unlawful means is punishable by five years' imprisonment and a fine of 300,000 euros".

[71] See for instance Article L. 2321-4 of the French 'Code de la défense' (defence code)

[72] Geoffroy Goubin, Lisa Janaszewicz, "Le *hacking* éthique : votre meilleur ennemi ?", *Dalloz IP/IT*, 2021 : 505

[73] See above, part II.

[74] Cambridge Analytica was a firm aiming at processing American citizens' data for political targeting purposes. However, the firm also processed data belonging to European citizens. From this background the Italian and British DPAs investigated on violations of the GDPR and the British data protection law, see GPDP, "Provvedimento del 10 gennaio 2019 [9080914] », January 10th, 2019, available at: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9080914,
and ICO, "Investigation into the use of data analytics in political campaigns : A report to Parliament", November, 6th, 2018 , p. 36, available at: https://ico.org.uk/media/action-weve-taken/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf, last accessed on 25th January, 2023

- The need to act in compliance with data subjects' privacy

When researchers carry out re-identification attacks, they do so to reveal vulnerabilities and to reinforce privacy. These attacks should not undermine privacy online. For instance, when a dataset has been re-identified, the researcher, even in the absence of any guidelines from the EDPB, should contact a data protection authority to ensure that such information does not fall into the wrong hands. It goes without saying, but de-anonymised data should not be made public.

3) Recommendations to companies and individuals releasing anonymised datasets

- The need to ensure a proper level of data confidentiality

In order to ensure a proper level of data confidentiality and be in line with the requirements of the GDPR, data controllers might ask for the help of their DPAs. As the Working Party 29 has stated, "*case studies and research publications have shown how difficult it is to create a truly anonymous dataset whilst retaining as much of the underlying information as required for the task*"[75].' In the same line of thought, the CNIL claimed that "*(g)iven the complexity of the issues involved in the choice and regular evaluation of anonymisation techniques, it is recommended that public authorities work on these in a concerted manner, in association with their data protection officers (...). The CNIL could also provide its expertise on the most frequently encountered problems on this subject (...)*".[76]
The same goes for some provisions of the French Law on data processing, files and liberties, which encourages researchers to request the CNIL's assistance[77].
Considering the difficulty of anonymisation, data controllers are thus strongly advised to contact their DPA to ensure a high degree of trust concerning their anonymisation technique. To summarise, they should adhere to the principle that prevention is better than cure.

Conclusion:

In this paper we first addressed the issue of anonymisation which is more and more difficult to implement with regard to the improvement of re-identification techniques such as inference or linkage. In the second part we came to the conclusion that there are still many grey zones in European data protection law when considering the status of anonymised data and the respective responsibility of the entity which anonymises data on the one hand, and the researcher carrying out re-identification for scientific purposes, on the other. Eventually we identified some recommendations in order to avoid any legal risks as regards re-identification, be it from data subjects' perspective or from the researchers' one. In particular, there is a pressing need for the EDPB and local data protection authorities to issue guidelines on this topic.

[75] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, last accessed on 25th January, 2023

[76] CNIL, « L'anonymisation des données, un traitement clé pour l'open data », October, 17th, 2019, available at : https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data, last accessed on 25th January, 2023

[77] Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, available at : https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460, last accessed on 25th January, 2023

Finally, this paper questions in a broader manner the relevance of Recital 26 of the GDPR. Since anonymisation cannot be fully reliable, anonymised data should always be subject to GDPR application.

Bibliography

Official sources:

European Union:

Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation (WP 203, 2013), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf, last accessed on January 25th, 2023.

Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, last accessed on January 25th, 2023.

ENISA, "Data Protection Engineering : From Theory to Practice'', January 2022, available at: https://www.enisa.europa.eu/publications/data-protection-engineering, last accessed on January 25th, 2023.

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

Data protection authorities' Opinions and Guidelines:

CNIL, « L'anonymisation des données, un traitement clé pour l'open data », October 17th, 2019, available at : https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data, last accessed on January 25th, 2023.

CNIL, « Recherche scientifique (hors santé) : enjeux et avantages de l'anonymisation et de la pseudonymisation », January, 31st, 2022, available at : https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation, last accessed on January 25th, 2023.

CNIL, « Régime juridique applicable aux traitements poursuivant une finalité de recherche scientifique (hors santé) » : 1-7, available at : https://www.cnil.fr/sites/default/files/atoms/files/consultation_publique_-_presentation_du_regime_juridique_applicable_aux_traitements_a_des_fins_de_recherche.pdf, last accessed on January 25th, 2023.

Datatilsynet, "The anonymisation of personal data", available at: https://www.datatilsynet.no/en/regulations-and-tools/reports-on-specific-subjects/anonymisation/?print=true, last accessed on January 25th, 2023.

GPDP, "Provvedimento del 10 gennaio 2019 [9080914] », January 10th, 2019, available at: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/908091, last accessed on January 25th, 2023.

ICO, “Investigation into the use of data analytics in political campaigns : A report to Parliament”, November, 6th, 2018 , p. 36, available at: https://ico.org.uk/media/action-weve-taken/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf, last accessed on January 25th, 2023.

Courts' decisions:

Jane Doe v. Netflix, Inc. et al, US District Court of the Northern District of Indiana, available at https://www.wired.com/images_blogs/threatlevel/2009/12/doe-v-netflix.pdf, last accessed on January 25th, 2023.

Doctrinal references:

Book:

Bygrave, Lee.A, *Data Protection Law: Approaching Its Rationale, Logic and Limits*, Kluwer Law International, 2002.

Articles:

Arrington Michael (August 6, 2006). "AOL proudly releases massive amounts of user search data". *TechCrunch*. Archived from the original on August 12, 2006. Retrieved August 7, 2006

Bannelier Karine, Anaïs Trotry, “What is ‘data’? Definitions in International Legal Instruments on Data Protection, Cross-Border Access to Data & Electronic Evidence”, *ai-regulation.com*, January 10th, 2023.

Barbaro Michael, Zeller Jr Tom Zeller “A Face Is Exposed for AOL Searcher No. 4417749”, *The New York Times*, August 9, 2006, available at : https://www.nytimes.com/2006/08/09/technology/09aol.html, last accessed on January 25th, 2023.

Cohen Aloni, “Attacks on deidentification's defenses”. In *31st USENIX Security Symposium* (USENIX Security 22), Boston, MA, (August 2022) : 1469-1486.

Dhondt, Karel, Victor Le Pochat, Alexios Voulimeneas, Wouter Joosen, and Stijn Volckaert. “A Run a Day Won't Keep the Hacker Away: Inference Attacks on Endpoint Privacy Zones in Fitness Tracking Social Networks” In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 801-814. 2022.

Dwork, Cynthia, Adam Smith, Thomas Steinke, Jonathan Ullman, “Exposed! A Survey of Attacks on Private Data.” *Annual Review of Statistics and Its Application,* (2017).

Finck, Michèle, and Frank Pallas. 2020. “They who must not be identified—distinguishing personal from non-personal data under the GDPR.” *International Data Privacy Law* 10 n°1, (2020) : 11–36.

Gambs, Sébastien, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. "De-anonymization attack on geolocated data." *Journal of Computer and System Sciences* 80, no. 8 (2014): 1597-1614.

Goubin Geoffroy, Lisa Janaszewicz, “Le ‘*hacking’* éthique : votre meilleur ennemi ?”, D*alloz IP/IT*, 2021 : 505

Helminger Lukas, Christian Rechberger, “Multi-Party computation in the GDPR”*, IACR Cryptol. ePrint Arch. 2022,* (2022) : 491

Maldoff Gabe, “How GDPR changes the rules for research”, *IAPP,* April 19th, 2016, available at: https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research/

Narayanan, Arvind, and Vitaly Shmatikov. "How to break anonymity of the netflix prize dataset." *arXiv preprint cs/0610105* (2006).

Ohm Paul, “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization.” *Ucla L. Rev.* n°57 (2009) : 1701.

Spindler Gerald, Philipp Schmechel, “Personal data and encryption in the European general data protection regulation” *J. Intell. Prop. Info. Tech. & Elec. Com. L.* n°7 (2016) : 163.

Stadler, Theresa, Bristena Oprisanu, and Carmela Troncoso. "Synthetic data–anonymisation groundhog day." In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1451-1468. 2022.

Sweeney Latanya , “k-Anonymity: A Model for Protecting Privacy”, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, n°10 (2002): 557-570

Sweeney Latanya, “Simple Demographics Often Identify People Uniquely”. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000, 1-34

Vokinger Kerstin Noëlle, Urs Jakob Mühlematter, “Re-Identifikation von Gerichtsurteilen durch ‘Linkage’ von Datenbanken”, *in : Jusletter,* 2 septembre 2019.

Willemson, Jan. "Fifty Shades of Personal Data–Partial Re-identification and GDPR." In *Privacy Technologies and Policy: 10th Annual Privacy Forum, APF 2022, Warsaw, Poland, June 23–24, 2022, Proceedings*, pp. 88-96. Cham: Springer International Publishing, 2022.

These statements are attributable only to the author, and their publication here does not necessarily reflect the view of the other members of the AI- Regulation Chair or any partner organizations.

This work has been partially supported by MIAI @ Grenoble Alpes, (ANR-19-P3IA-0003)

This work has been supported by the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR.

This work was co-funded by the Cybersecurity Institute of Grenoble Alpes (ANR-15-IDEX-02)