Anonymisation and re-identification risk for voice data

Alvaro Moretón, Ariadna Jaramillo

Practitioners' Corner

Anonymisation and Re-Identification Risk for Voice Data

*Alvaro Moretón and Ariadna Jaramillo**

I. Introduction

This document present several interpretations of the concept of anonymisation provided in the GDPR, including the requirements for effective anonymisation. It focuses on the automatic anonymisation of voice data in voice assistants and voice-enabled applications and the issues that may arise from it, particularly the re-identification risk of data subjects and the evaluation of such risk. It relies on H2020 project COMPRISE to further explain the different issues and possible solutions to reach anonymisation of voice data in voice-enabled systems.

Anonymisation is addressed in the General Data Protection Regulation (GDPR)[1] as a protective measure to prevent natural persons from being identified or becoming identifiable, which means that anonymised data is no longer personal and therefore no longer falls within the scope of the GDPR (Recital 26).

This report analyses different requirements for effective anonymisation following the GDPR, research work and guidelines issued by the Article 29 Working Party (now succeeded by the EDPB) and national supervisory authorities such as the UK Information Commissioner's Office (ICO), the Irish Data Protection Commission (DPC), the Spanish Agencia Española de Protección de Datos (AEPD) and the French Commission Nationale Informatique et Libertés (CNIL), and it compares their notions of anonymisation. More specifically, it focuses on the anonymisation of voice data in voice assistants and voice-enabled applications and the issues that may arise from it, particularly the re-identification risk of data subjects and the evaluation of such risk.[2] It delves on factors that impact re-identification (e.g., the user's profile and context), possible means of re-identification (e.g., inference), and proposes measures to mitigate the re-identification risk (e.g., data governance measures).

The H2020 project COMPRISE (Cost-effective, Multilingual, Privacy-driven voice-enabled Services)[3] has introduced the first voice and text anonymisation methods designed specifically to protect the privacy of voice interface users. In the following sections, we rely on these methods to illustrate the issues mentioned above and discuss the benefits and limitations of future privacy-driven voice assistants and voice-enabled applications.

II. The Concept of Anonymisation

This section analyses Recital 26 of the GDPR, which provides a first notion on the concept of anonymisation. Likewise, it analyses the interpretation that various supervisory authorities and the former Article 29 Working Party (now succeeded by the European Data Protection Board) have made on the concept of anonymisation provided in the Regulation.

1. Anonymisation in the GDPR: Recital 26

The General Data Protection Regulation (GDPR) firsts refer to anonymisation in Recital 26 '[...]To determine

DOI: 10.21552/edpl/2021/2/20

* Alvaro Moretón is the corresponding author and Project manager at Rooter Analysis SL in Madrid; for correspondence <alvaro.moreton@rooter.es>; Ariadna Jaramillo is project intern with Rooter; for correspondence: <ariadna.jaramillo@rooter.es>. The work described in this report was partly supported by the European Union's Horizon 2020 Research and Innovation Program, under Grant Agreement No. 825081 COMPRISE <https://www.compriseh2020.eu/>. Emmanuel Vincent (Senior Research Scientist and Project Coordinator) and Marc Tomassi (Professor in Computer Science at Lille University), both part of the COMPRISE project, contributed to this article by providing comments and feedback in their areas of expertise.

1 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119 (2016) 1–88.

2 On the question of identification of personal data in such systems cf. Moreton/Jaramillo, 'How can Private Information Recorded by Voice-enabled Systems be Identified?' (2020) 6 EDPL 3 464 – 469.

3 See, <https://www.compriseh2020.eu/>

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. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable'.

It follows from the above that a dataset is deemed anonymised when the individuals behind can no longer be identified or are no longer identifiable, given that it is not possible to individualise them through different pieces of information. Therefore, if a dataset has been effectively anonymised, it is not considered personal data, and the GDPR does not apply to it.

2. Interpretations of Recital 26

From the concept provided in Recital 26, it appears that the GDPR embraces a risk-based approach to qualifying data either as personal or non-personal (i.e., anonymised). In this sense, whenever the risk of re-identification is considered reasonable, the data is deemed personal. In contrast, if the risk of re-identification is negligible, the data is considered non-personal, in which case the GDPR does not apply.[4]

The main problem with anonymised datasets is to ensure that a sufficient level of anonymisation has been reached so that no individual can be re-identified. This comes with the additional complication of maintaining the utility of the anonymised dataset for the desired task.

a. The Article 29 Working Party

According to the Article 29 Working Party (Article 29 WP), Recital 26 of the GDPR should be interpreted more stringently: personal data should be considered anonymised and consequently not subject to the GDPR only when anonymisation is irreversible.[5] Such an interpretation diverges from the spirit of Recital 26, which states that anonymisation should be assessed based on the risk of re-identification and accepts a 'tolerable risk'. This divergence creates uncertainty when assessing anonymisation, since the expression 'irreversible process' employed by the Article 29 WP seems to imply that no remaining risk of re-identification is tolerable for data to be deemed anonymised.[6]

b. Supervisory Authorities

Supervisory authorities have also taken positions on the requirements for anonymisation. Most of them favour a balance between the risk-based approach of Recital 26 and the strict interpretation of the Article 29 WP.

For instance, the UK's ICO acknowledges that 'the risk of re-identification through data linkage is essentially unpredictable because it can never be asserted with certainty what data is already available or what data may be released in the future'.[7] Therefore, the relevant criterion to consider data as personal is the 'identification or likely identification of the data subject'. The Irish DPC took a similar approach by considering that it is enough to demonstrate that re-identification is highly unlikely given the specific circumstances.[8]

Conversely, the French CNIL embraces a strict interpretation of Recital 26 by considering that anonymisation is only achieved when 'identification is practically impossible'. However, the authority clarifies that the impossibility of identification is the goal when stating that anonymisation 'seeks to be irreversible'.[9]

4 Michèle Finck, Frank Pallas, 'They who must not be identified—distinguishing personal from non-personal data under the GDPR' (2020) 10 International Data Privacy Law 1.

5 Article 29 WP, Opinion 05/2014 on Anonymisation Techniques, adopted on 10 April 2014, WP 216, 5.

6 (n 4).

7 Information Commissioner's Office, 'Anonymisation: managing data protection risk code of practice' (November 2012) <https://ico.org.uk/media/1061/anonymisation-code.pdf> accessed 27 June 2021.

8 Data Protection Commission, 'Guidance on Anonymisation and Pseudonymisation' (June 2019) <https://www.dataprotection.ie/sites/default/files/uploads/2020-09/190614%20Anonymisation%20and%20Pseudonymisation.pdf> accessed 27 June 2021.

9 (n 4).

Considering all the opinions above, and as it seems that achieving a total, irreversible anonymisation is almost impossible due to the lack of a manageable technique that guarantees a 100 % irreversibility and the total inexistence of the re-identification risk, the most reasonable and realistic way to deal with anonymisation appears to be through a risk-based approach that determines whether data qualifies as personal or non-personal after applying the corresponding technique to the dataset. This approach seems to align better with the basic idea of the GDPR, which encourages the risk-based approach when implementing the adequate organisational and technical measures necessary to comply with the Regulation.

III. Anonymisation of Voice Data

Assessing the risk of data subject's re-identification is a complex task, especially when anonymisation is applied to the data collected through voice assistants or voice-enabled applications that process massive amounts of data (due to the constant man-machine interaction) from a very large number of users. We illustrate this issue through an analysis of the COMPRISE solution.

COMPRISE (Cost-effective, Multilingual, Privacy-driven voice-enabled Services) is an H2020 project funded by the European Commission that has introduced a complete methodology to ensure the privacy of voice interface users. Unlike most voice technology providers that rely on cloud-based speech-to-text and spoken language understanding components to transcribe the user's voice input into text and infer the user's intent, COMPRISE proposes to run these two components on the user's device or a trusted server. While this preserves the user's privacy at inference time, i.e., when using existing speech-to-text and spoken language understanding components, the technology provider must still store speech and text data in the cloud and manually label it to train the AI models behind these two components and improve them over time.

COMPRISE ensures privacy at training time by automatically anonymising speech and text data before sending them to the cloud via the following tools:

– The privacy-driven voice transformer: This tool is applied to the user's voice signal. It results in a new de-identified voice signal from which the user's identity has been removed while the remaining information is ideally unchanged. In addition, words and expressions carrying personal information are deleted from the voice signal. Only a short-duration portion (in the order of 2 s for every input utterance) is kept. The anonymised voice data gathered from all users can then be labelled and used to train a speech-to-text model.
– The privacy-driven text transformer: This tool is applied to the textual transcription of the user's voice signal (i.e., the speech-to-text output). It identifies words and expressions carrying personal information and replaces them with random alternatives while preserving the sentence structure. The anonymised text data gathered from all users can then be labelled and used to train a spoken language understanding model.

For a given utterance, either the voice signal or the corresponding text are sent to the Cloud. Indeed, transmitting the de-identified text (where words and expressions carrying personal information have been replaced) and the de-identified voice signal (where words and expressions carrying personal information have been deleted) at the same time would reveal those words and expressions carrying personal information, which have not been identified by the text transformer and can, therefore, help re-identify the user. Also, the data is transmitted without any additional information (e.g., IP address, pseudonymous identifier, etc.), making it possible to link several voice signals or pieces of text with each other and increase the chances of re-identification. Finally, speech-to-text and spoken language understanding models are trained using weakly supervised learning by labelling the de-identified voice and text data automatically by means of automatic weak labellers. This helps reduce the potential risk of privacy breaches by reducing the amount of data that need human labelling.

Considering the above, it seems logical that automatic anonymisation affects data controllers' tasks, mainly because they will not have complete control over the data that is anonymised (as anonymisation is automatised) and directly uploaded into the cloud, and the risk of re-identification of the dataset will depend on the specific circumstances related to each dataset. This means that an automatic anonymisation system may be more or less effective depending on the characteristics of each dataset (i.e., the content of the information contained in the dataset or the cir-

cumstances of the individual from which such information is obtained), making it difficult to achieve a fully individualised analysis.

The following aspects may influence anonymisation effectiveness, and consequently, the risk of re-identification of an individual:

- Specific circumstances of the data subject (e.g., how much information about the individual is available for linkage purposes) as contextual aspects;
- Exhaustive definitions of the words and expressions that should be suppressed or masked as for example, the data subject could be easily identified through a phone number. Suppose the anonymisation tool has been programmed to detect specific words and expressions containing the subject's name, gender, or health condition only. In that case, telephone numbers will remain in a dataset, and the data subject could be re-identified through them;
- Correct automatic detection of these words and expressions as accurate detection may depend on aspects such as speech-to-text errors, pronunciation, trigger phrases or way of speaking, e.g., use of colloquial language or slang.

IV. Aspects to be Considered when Assessing Re-Identification Risks

This section explores the concepts of identifier and quasi-identifier and their role in the possible re-identification of the data subject. Means of re-identification, such as singling and linkage, are analysed theoretically and through use cases involving automatic anonymisation and COMPRISE.

1. Identifiers and Quasi-Identifiers

According to the ICO, the concept of 'identifiable or anonymisation' is not completely clear, as individuals can be identified differently.[10] For instance, individuals can be directly identified through a single data source, also known as direct identifiers. Below are listed a few examples[11]:

- Unique**:** Created for specific administrative purpose and associated directly with an individual (e.g., personal ID, social security number);
- Associational**:** Labels or data related to objects with strong and enduring associations with individuals (e.g., mobile phones, vehicle registration);
- Transactional**:** Labels or data strings associated with individuals within the scope of a particular transaction (e.g., dynamic IP addresses, cookies or email alias).

Individuals could also be identified through different pieces of information called quasi-identifiers, defined as a set of attributes that can be used to identify a person. In this scenario, re-identification can be achieved either by using quasi-identifiers from a single dataset or by combining quasi-identifiers from different sets. Below are listed some examples of quasi-identifiers[12]:

- Physical attributes;
- Race and ethnicity;
- Profession;
- Gender.

Quasi-identifiers play an important role in anonymisation processes. Recent research shows that only a few are enough to re-identify an individual. For instance, 1% of the Swedish population can be identified only through their age, occupation, municipality and gender, all quasi-identifiers found in Statistiska Centralbyrån (Central Bureau of Statistics) tables[13]. Following the same line, researchers from Imperial College London and Belgium's Université Catholique de Louvain have recently published a method they claim is able to re-identify 99.8% of individuals in anonymised datasets using less than 15 demographic results. According to the results, the more attributes in the dataset, the more likely the re-identification. Nonetheless, even with few attributes available, re-identification is still possible.[14]

The re-identification risk largely depends on the effectiveness of the anonymisation technique em-

10 Information Commissioner's Office, 'Anonymisation: managing data protection risk code of practice', Information Commissioner's Office, November 2012, <https://ico.org.uk/media/1061/anonymisation-code.pdf> accessed 27 June 2021.

11 Mark James Elliot et al, 'Functional anonymisation: Personal data and the data environment' (2018) Computer Law & Security Review.

12 Electronic Health Information Laboratory, 'What is a quasi-identifier?' (Unknown) <http://www.ehealthinformation.ca/faq/quasi-identifier/> accessed 27 June 2021.

13 Nyhet Verksamhetsskydd N., 'Quasi identifiers and the challenges of anonymising data' (Basalt, 30 January 2017) <https://www.basalt.se/news/quasi-identifiers-and-the-challenges-of-anonymising-data/>

14 Luc Rocher et al, 'Estimating the success of re-identifications in incomplete datasets using generative models' (2019) Nature Communications.

ployed. In this regard, formal anonymisation, the procedure through which direct identifiers are cleared from a dataset, might not be effective enough to avoid re-identification but sufficient to block a direct re-identification by an adversary (which partially fulfils the GDPR requirements). However, re-identification could still be possible using quasi-identifiers that have not been removed.[15]

Therefore, formal anonymisation is inadequate if a low risk of re-identification is desired and should be considered a minimal intervention. Unfortunately, more intensive anonymisation techniques could be complex to achieve without compromising the practical utility of the dataset.

Considering the above, a viable solution to complement weak anonymisation (in which case GDPR would still apply to the processing operations) could be implementing additional controls and data governance measures. Even when anonymisation effectively minimises personal data processing (complying with the minimisation principle on Article 5 (1) (c) GDPR), the uncertainty and potential risk of re-identification might be mitigated by applying additional minimisation measures.[16]

The risks arising from employing an automatic anonymisation system (e.g., the system fails to recognise identifiers or quasi-identifiers as personal data) are analysed for COMPRISE as well, which can be explained considering the example that a company has developed a dictation-based app that allows students to record notes and reminders on curricular and extracurricular activities (e.g., tournaments, rehearsals, practices, examinations, etc.). Though COMPRISE tools neutralise the users' voice and anonymise data related to locations, addresses, dates, names, etc., some students use slangs or acronyms to address personal data, preventing the app to correctly identify them as such. Even the student's pronunciation could lead to speech-to-text errors and prevent COMPRISE tools from recognising a specific piece of information as personal data.

Therefore, to mitigate the risk of re-identification if the automatic identification of private words that should be anonymised fails (or the system is configured to detect only specific categories of private information), COMPRISE follows an approach that allows exchanging critical words with different words of the same type. For instance, if the speaker utters the word 'May' in a phrase, it could be replaced by a different month selected randomly (e.g., 'June', 'October'). This way, even when the identification of private information fails to detect a relevant word occurrence, it would not be easy for an attacker to distinguish whether the words in the transformed transcript are the result of an actual transformation or the words initially uttered by the speaker.

The possibility of finding more or fewer identifiers and/or quasi-identifiers will depend on various aspects, including:

- The type of voice app through which the user's voice is collected: Predicting which personal information may be revealed by the speaker will be complex on apps that enable more open and unrestricted interactions.
- The length of the datasets collected by the app: The longer the dataset, the more information related to a single interaction it will contain.

To mitigate re-identification risk related to the length of the datasets collected through voice apps and minimise the number of possible identifiers and/or quasi-identifiers in these datasets, COMPRISE is assessing an anonymisation approach in which the length of the utterances is reduced, cutting them into small pieces of 1.5 to 2 s duration (note: a person's word rate is around 300/min., so 2s can be 10 words).

However, if the shortened utterance contains identifiers like names/surnames (e.g., Eve Williams) or ID numbers, said information would be considered personal data.

In the case of re-identification through quasi-identifiers, the probability of re-identification seems to reduce considerably with utterances of around 5-10 words in the context of normal interaction with a voice technology system. However, it is still possible, especially if all 10 words in the dataset are quasi-identifiers (e.g., 'Head of Legal at Bank X, Paris'). This probability would increase considerably if the dataset were linked to other external sources of information. In addition to the foregoing, there would always be some degree of uncertainty given that it is difficult to address whether the sentence is about the individual (i.e., Head of Legal) or said by this individual.

15 (n 11).

16 Information Commissioner's Office, 'Anonymisation: managing data protection risk code of practice', Information Commissioner's Office, November 2012, <https://ico.org.uk/media/1061/anonymisation-code.pdf> accessed 27 June 2021.

The preceding, combined with COMPRISE word's substitution function, is what is going to decrease re-identification risk significantly. In an ideal scenario, each developer or user would be able to select their own level of privacy based on utility, i.e. also determine whether anonymisation takes place or not.

Nonetheless, although it is now known by COMPRISE that the shortening of the datasets would have a very limited impact on utility, the possibility an attacker could put together the pieces of information to guess the initial sentence is yet uncertain. However, it should not be overlooked that this kind of attack cannot be measured by the degree of success or failure the attacker has. Still, it indeed contributes to reduce uncertainty and increase the re-identification risk.

2. Means of Re-Identification

There are several methods adversaries can employ to re-identify individuals in a dataset. Below are explained some of the most relevant.

a. Inference

The Article 29 WP defines inferences as 'the possibility to deduce, with significant probability, the value of an attribute from the values of a set of attributes'.[17] This definition opens further questions on the meaning of 'significant', particularly which level of probability is enough to consider that an inference has become a re-identification.

Below are listed some statements and commands from which it would be possible to infer an individual's identity or additional information concerning its persona in a speech-to-text system, even after automatic anonymisation (on specific words detected by the tool) has been applied:

- A voice-based app to dictate delivery orders — 'Deliver the book 'How to overcome depression' to Anselmo Fuentes at ByBob company, Picasso building, 3rd Floor, Office 3': Although the information highlighted bold has been anonymised, it would still be possible for an attacker to re-identify Anselmo simply by linking any publicly available information on his persona (e.g., his office address). It would also be possible for an attacker to infer information on Anselmo's mental health through the book's name, which hasn't been neutralised for not being considered personal or sensitive data by the system, though it suggests a possible mental condition (i.e., depression).
- A voice-based app for classifying and locating information more efficiently in legal environments — 'Find the documents: Case: Jones Brothers Corporation, Client: Jimmy Jones, Consultation: Criminal Liability of CEOs': Although the information highlighted bold has been anonymised, it would still be possible for an attacker to re-identify the client's identity and infer that his persona is involved in a legal case. Also, additional information such as the company's name to which the consultation is related could be inferred from the case name.

b. Linkage

The Article 29 WP defines linkability as 'the ability to link, at least, two records concerning the same data subject or a group of data subjects'.[18] The technique in question relies on the linkage of variables that are present in different datasets.[19] It can be performed by an intruder using personal data it already possesses (e.g., information available on a personal database) and matching it with information from an anonymised dataset, or using information from an anonymised dataset and trying to match it with available external information (e.g., information available on the internet).

Linkable information, therefore, can be both information available only to certain organisations or individuals or publicly available information accessible to virtually anyone.

The risk of re-identification through linkage will rise as new techniques are developed, computing power increases, and more data becomes publicly available. Nonetheless, quasi-identifiers also play an important role in linkage attacks. In this sense, even when direct identifiers have been suppressed from a dataset, it could still be relatively easy for an intruder to identify an individual through quasi-iden-

17 WP216 (n 5), 12.

18 ibid 11.

19 Government Data Quality Hub, 'Privacy and data confidentiality methods: a Data and Analysis Method Review (DAMR)' (*Government Statistical Service*, 13 December 2018) <https://gss.civilservice.gov.uk/wp-content/uploads/2018/12/11-12-18_FINAL_Kerina_Jones_David_Ford_article.pdf> accessed 27 June 2021.

tifiers, especially if combined with additional information.

Hence, re-identification through quasi-identifiers (that have not been removed from the datasets) is one of the main risks to be considered in systems employing automatic anonymisation (e.g., COMPRISE), given that focus is regularly placed on a few specific identifiers and quasi-identifiers. The best solution to tackle this issue is for data controllers to broaden to the extent possible (always considering the impact of excessive privacy on the data's utility) the number of quasi-identifiers that must be deleted or substituted.

i. Publicly Accessible Data

Public data (e.g., data contained in newspapers, social networks, blogs, etc.) could serve to re-identify individuals through linkage by combining it with data in anonymised datasets.

For instance, in 2006, Netflix had to cancel a competition due to privacy-related concerns linked to de-anonymisation. The researchers participating in the contest were competing to provide the best improvement for Netflix's suggestion algorithm, for which the platform published an anonymised sample of its movie ratings consisting of 10 million ratings from half a million users. However, a group of researchers from the University of Texas managed to de-anonymise users 68% of the time by correlating their private Netflix ratings with their public ratings on the Internet Movie Database (IMDB) and this research was conducted nearly fifteen years ago.[20]

ii. Data available to the Organisation

As noted in the subsection above, linkage can be performed by combining anonymised datasets with other information that only specific individuals or organisations can access. It could be the case of a recruitment company that employs its own team of developers to design a voice app to search for job openings. Even if all personal data concerning the user's professional profile and contact details is anonymised before it is sent to the cloud platform, the original outcome is still sent to the service provider (as it needs these pieces of information to provide the service), which happens to be the same company that designed the app.

According to the Article 29 WP, when a data controller does not delete the original, identifiable data from a dataset, and hands over part of it (for instance, after anonymising the dataset) to a third party, the resulting dataset will still be considered to be personal data. Hence, as long as the data controller can access the original raw data, even if direct identifiers have been removed from the set provided to the third party, the resulting dataset is deemed as personal data. However, if the data controller deletes the raw data and only provides aggregate statistics to third parties on a higher level, it would qualify as anonymised data. [21]

Additionally, organisations developing apps may have access to different personal information collected from their different apps, which might be necessary for it to run properly. So, even when some datasets collected by the app are anonymised (such as the voice data collected through the voice-enabled application, as in the case of COMPRISE), others (containing personal information) could be combined with the anonymised datasets, enabling the re-identification of the individual in the anonymised dataset, hence becoming personal data as well. Below are listed some examples of other sources of personal information that may be collected through apps:

- Access to the user's calendar: Voice apps usually access calendars to manage calendar entries via commands. However, the data contained in the calendar app may be sensitive in nature (e.g., an appointment with a specialist).
- Location data: Voice apps usually collect location data to provide accurate answers to the user's command (e.g., 'Show me the closest Chinese restaurant').
- Metadata collected through Installed Application Methods (IAMs): IAMs are tools used by several apps to collect information to find incompatibilities between apps. The IAMs collect information, such as lists of apps installed on a device and other metadata used to infer personal information about the user. Given that IAMs do not require specific users' permissions, many developers use them to collect such information.

---

20 Arvind Narayanan,Vitaly Shmatikov, 'Robust De-anonymization of Large Sparse Datasets' (2007) <https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf> accessed 27 June 2021.

21 Dan Cooper, Kristof Van Quathem, 'European Regulators Set Out Data Anonymisation Standards', Inside EU Life Sciences (2014) <https://www.insideeulifesciences.com/2014/04/24/european-regulators-set-out-data-Anonymisation-standards/> accessed 27 June 2021.

3. Data Subject's Profile

Another aspect that should be considered when assessing the re-identification risk is the data subject's profile, whose data has been anonymised and the impact and damage to which he/she would be exposed in the case of a potential re-identification.

On the one hand, an intruder/attacker's motivation could depend on the user's profile. For instance, an economically privileged individual could be targeted for financial benefits (e.g., attacks aiming at blackmailing him/her, obtain passwords, or create a fake identity). Furthermore, depending on the data subject's profile, more or less information on his/her persona would be publicly available. In principle, public figures (e.g., artists, politicians, professionals with public notoriety like a lawyer or a doctor, etc.) would be considerably more vulnerable as more public information about them is available on the Internet (e.g., magazines or newspapers articles, videos, blogs etc.), making easier the linkage of datasets and re-identification.

Additionally, the negative impacts derived from disclosing private information of public figures could be higher than those of the general population, as the information leaked about a public figure, if it is sensitive, probably will be published in media and may damage its reputation (as a consequence, its personal and professional life may result highly affected) . In these cases, the ICO[22] recommends adopting a more rigorous form of risk analysis and anonymisation to inform the data subject on possible consequences derived from the disclosure and to seek his/her consent.

In COMPRISE, the main problem of voice-enabled apps and the datasets uploaded to the COMPRISE platform is that anyone can be an app user, especially in generalist apps that allow virtually any user profile. It is not possible to know whether the anonymised voice and text data come from a high-profile individual (i.e., a person known to the public/with immediate public visibility which has significant amounts of personal information available to the public) or someone unknown to the general public, and therefore to identify which individuals will be more affected by a potential re-identification and/or data disclosure.

Moreover, as the anonymisation is carried out automatically and applied to all data indistinctly, there would be no intervention from the data controller to evaluate the user's profile and possible re-identification risks before anonymising the data and storing it in the platform.

V. Data Governance

It is important to implement an adequate data governance framework to evaluate the effectiveness of the anonymisation techniques and additional measures when required.

1. Assessment of the Anonymisation

Depending on the effectiveness of the anonymisation techniques implemented, additional measures might be needed in order to lower re-identification risks to tolerable levels. However, at one point, it is possible that additional measures will not add significant safeguards but, on the contrary, significantly reduce the data value due to privacy protectionism.[23]

The aspects discussed in Section III play a role in assessing whether the re-identification risk is tolerable or unacceptable.

The ICO introduced the 'motivated intruder test' as a good practice to assess the re-identification risk in anonymised datasets. According to its guidelines[24], the motivated intruder test consists of a search through different, accessible information sources to find information that, in combination with the anonymised dataset, may result in the individual's re-identification. However, it might struggle with systems where anonymisation is automatic (automatic anonymisation and automatic upload to the Cloud), like COMPRISE.

In COMPRISE, as the same anonymisation technique is employed for all the datasets obtained through multiple voice-enabled apps, the motivated

22 Information Commissioner's Office, 'Anonymisation: managing data protection risk code of practice', Information Commissioner's Office (2012) <https://ico.org.uk/media/1061/anonymisation-code.pdf>.

23 Government Data Quality Hub, 'Privacy and data confidentiality methods: a Data and Analysis Method Review (DAMR)', Government Statistical Service (2018) <https://gss.civilservice.gov.uk/wp-content/uploads/2018/12/11-12-18_FINAL_Kerina_Jones_David_Ford_article.pdf> accessed 27 June 2021.

24 Information Commissioner's Office, 'Anonymisation: managing data protection risk code of practice' (*Information Commissioner's Office*, November 2012) <https://ico.org.uk/media/1061/anonymisation-code.pdf > accessed 27 June 2021.

intruder may be successful in some cases and unsuccessful in others, depending primarily on factors such as the content of the dataset, or the information available on the user, to mention a few.

Furthermore, if the re-identification risk has been assessed as intolerable and the anonymisation effectiveness is uncertain, the wisest solution would be to consider the anonymised dataset as personal data. That way, all the organisational and technical measures to comply with the GDPR requirements and protect personal data in the dataset would be applied.

Even when the data controller is incapable of identifying an individual within a dataset that has been automatically anonymised only using the information in it, there is still a latent risk of re-identification through linkage. In a scenario like this, the controller could opt for managing the dataset as personal data and hence (must) comply with the GDPR requirements. But how will the data controller fulfil its obligations regarding data subjects' rights if it cannot identify them in the 'anonymised' datasets? (e.g., how will the exercise of the right to access by the data subject be enabled if it is not possible to identify him/her within the dataset?)

Article 11 of the GDPR provides a possible solution for these cases:

1. If the purposes for which a controller processes personal data do not or do no longer require the identification of a data subject by the controller, the controller shall not be obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with this Regulation.
2. Where, in cases referred to in paragraph 1 of this Article, the controller is able to demonstrate that it is not in a position to identify the data subject, the controller shall inform the data subject accordingly, if possible. In such cases, Articles 15 to 20 shall not apply except where the data subject, for the purpose of exercising his or her rights under those articles, provides additional information enabling his or her identification.

According to Article 11, it becomes clear that, on one side, the GDPR should not be deemed an excuse for processing more data than strictly necessary and, on the other, that it serves as a tool to limit some of the data controller's obligations.[25]

Another requirement of the GDPR that involves the need to identify the data subject for its fulfilment is the obligation of communication of a personal data breach to the data subject when the data breach is likely to result in a high risk to the rights and freedoms of natural persons. However, Article 34 of the GDPR, which establishes this obligation, also states that the communication to the data subject 'shall not be required if any of the following conditions are met:

1. The controller has implemented appropriate technical and organisational protection measures, and those measures were applied to the personal data affected by the personal data breach, in particular those that render the personal data unintelligible to any person who is not authorised to access it, such as encryption.
2. The controller has taken subsequent measures which ensure that the high risk to the rights and freedoms of data subjects referred to in paragraph 1 is no longer likely to materialise.
3. It would involve a disproportionate effort. In such a case, there shall instead be a public communication or similar measure whereby the data subjects are informed in an equally effective manner.'

So, again, if the controller cannot identify the individual of an anonymised dataset, but there is still a considerable risk of re-identification by linkage, in the event that a data breach takes place, the data controller would be not required to communicate it to the data subject based on condition 3. However, it should still be necessary to issue a public communication or a similar measure.

Also, depending on the effectiveness of the anonymisation and the implementation of additional measures, condition 1 would be applicable if it's possible to demonstrate that attackers wouldn't be able to identify the data subject.

2. Other Data Governance Measures

Below are briefly described some additional governance measures that can be implemented to mitigate the risk of re-identification:

25 Alina Skiljic, 'Article 11 GDPR: Processing data that does not require identification and how it should not be interpreted', (The Privacy Advisor, 27 October 2020) <https://iapp.org/news/a/article-11-gdpr-processing-data-that-does-not-require-identification-and-how-it-should-not-be-interpreted/> accessed 27 June 2021.

a. Sharing Restrictions

Given the variety of anonymised data that can be derived from personal data, data controllers need to consider their disclosure options carefully (i.e., who they are sharing or disclosing the dataset with).[26] Depending on the disclosure level, controls could be more stringent or more flexible. For example, an open data environment leaves no residual element of control as it is very permissive. Hence, it demands a very secure derived dataset.[27]

To restrict the sharing and use of datasets, some requirements could be demanded to operate in the corresponding environment. For instance, adhering to an ethics code or certification issued by accredited certification bodies (e.g., ISO 27001). In the case of COMPRISE, annotators and developers using the solutions could be asked to adhere to these codes or to prove they have been certified by an accredited certification body.

Good practices could also serve to mitigate re-identification risks. For example, applying policies against any attempts to re-identify data subjects from the anonymised dataset and include in the contracts' clauses binding signatory parts (i.e., parties that will have access to the data) to professional and ethical obligations[28] allow developers to configure sharing restrictions and generate sharing policies of each app, etc.

b. Structural Governance Measures

The ICO recommends that organisations implement structural governance measures regardless of whether the datasets resulting from an anonymisation process are deemed as personal data or not. Such measures include procedures to identify cases where anonymisation may be problematic (e.g., difficulty to assess the re-identification risk), carrying out Privacy Impact Assessments to test the effectiveness of a given anonymisation technique and help assess and mitigate the re-identification risk, and/or implementing access control rules both for anonymised data and original data, to mention a few.

c. Actions to Ensure Unlinkability

The European Agency for Cybersecurity (ENISA) provides some recommendations to ensure the unlinkability of datasets that may result in the re-identification of individuals by app developers and/or app providers:

- For each app, only personal data necessary for the stated purpose should be processed;
- Different apps processing personal data for different purposes should be isolated by default and data exchange should be prevented unless explicitly specified or otherwise chosen by the user;
- The app's default configuration must ensure that only personal data necessary for the purpose of the processing is processed.

VI. Recommendations

Before presenting the conclusions to this report, we list a series of recommendations aimed at reducing the risk of re-identification in systems that employ automatic anonymisation:

- Since achieving complete, irreversible anonymisation is almost impossible due to the lack of a technique that guarantees a hundred per cent irreversibility and the total inexistence of the re-identification risk, it would be reasonable for data controllers to deal with anonymisation through a risk-based approach that determines whether data qualifies as personal or non-personal after applying the corresponding technique to the dataset.
- If the resulting risk remains intolerable after performing the re-identification risk assessment and the anonymisation effectiveness is uncertain, it is recommended to consider the anonymised dataset as personal data. This way, all organisational and technical measures in the GDPR to protect personal data must be applied (e.g., sharing restrictions, structural governance measures, etc.).
- It is recommended to implement additional controls and data governance measures to complement weak anonymisation. This would help miti-

26 Information Commissioner's Office, 'Anonymisation: managing data protection risk code of practice', Information Commissioner's Office (2012) <https://ico.org.uk/media/1061/anonymisation-code.pdf> accessed 27 June 2021.

27 (n 11).

28 International Association of Privacy Professionals, 'European Legal Requirements for Use of Anonymised Health Data for Research Purposes by a Data Controller with Access to the Original (Identified) Data Sets', Resource Center (2017) <https://iapp.org/resources/article/european-legal-requirements-for-use-of-anonymized-health-data-for-research-purposes-by-a-data-controller-with-access-to-the-original-identified-data-sets/> accessed 27 June 2021.

gate the uncertainty and potential risk of re-identification remaining after the corresponding anonymisation technique aimed at minimising personal data processing is applied.
- Data controllers of systems employing automatic anonymisation should broaden, to the extent possible (always considering the impact of excessive privacy on the data utility), the number of quasi-identifiers that must be deleted or substituted from the datasets to reduce the risk of linkage attacks.
- Data controllers should be aware, to the extent possible, of pieces of information concerning data subjects publicly available on the internet or to the organisation before releasing any anonymised dataset contained information on their personas, as it could be used for linkage.

VII. Conclusions

The risk of re-identification will never cease to exist, as achieving a total, irreversible anonymisation is virtually impossible without sacrificing the dataset utility. Therefore, it is recommended to perform a re-identification risk assessment whenever anonymisation is applied to the dataset, even for systems that employ automatic anonymisation and that automatically share the anonymised sets, independently of the technique's effectiveness, like COMPRISE. In this context, though there are different approaches on how to perform a re-identification risk assessment, the most realistic is to accept a tolerable risk of re-identification after anonymisation.

The assessment results should serve as a compass to decide whether additional measures should be implemented to safeguard the anonymised dataset, but especially if the dataset would be considered personal or non-personal data (i.e. is effectively anonymised), with all the implications in terms of compliance that the former implies. The preceding takes on greater relevance for systems using automatic anonymisation, where additional measures should be stringent.

The re-identification risk assessment considers several elements, such as possible identifiers and quasi-identifiers in the dataset, possible means of re-identification, sources where public information may be gathered, the data subject profile, etc. In the end, all of them will serve the data controller to decide on the optimal ways to safeguard, share or disclose the dataset, as well as to pinpoint weaknesses that should be corrected to improve anonymisation. The goal is to prevent the anonymised dataset from revealing the data subject's identity if combined with external information, for instance, through linkage.