

# Privacy in Speech and Language Technology

Simone Fischer-Hübner<sup>\*1</sup>, Dietrich Klakow<sup>\*2</sup>, Peggy Valcke<sup>\*3</sup>, and Emmanuel Vincent<sup>\*4</sup>

1 Karlstad University, SE. [simone.fischer-huebner@kau.se](mailto:simone.fischer-huebner@kau.se)

2 Saarland University – Saarbrücken, DE. [dietrich.klakow@lsv.uni-saarland.de](mailto:dietrich.klakow@lsv.uni-saarland.de)

3 KU Leuven, BE. [peggy.valcke@kuleuven.be](mailto:peggy.valcke@kuleuven.be)

4 Inria – Nancy, FR. [emmanuel.vincent@inria.fr](mailto:emmanuel.vincent@inria.fr)

---

## Abstract

This report documents the outcomes of Dagstuhl Seminar 22342 “Privacy in Speech and Language Technology”. The seminar brought together 27 attendees from 9 countries (Australia, Belgium, France, Germany, the Netherlands, Norway, Portugal, Sweden, and the USA) and 6 distinct disciplines (Speech Processing, Natural Language Processing, Privacy Enhancing Technologies, Machine Learning, Human Factors, and Law) in order to achieve a common understanding of the privacy threats raised by speech and language technology, as well as the existing solutions and the remaining issues in each discipline, and to draft an interdisciplinary roadmap towards solving those issues in the short or medium term.

To achieve these goals, the first day and the morning of the second day were devoted to 3-minute self-introductions by all participants intertwined with 6 tutorials to introduce the terminology, the problems faced, and the solutions brought in each of the 6 disciplines. We also made a list of use cases and identified 6 cross-disciplinary topics to be discussed. The remaining days involved working groups to discuss these 6 topics, collaborative writing sessions to report on the findings of the working groups, and wrap-up sessions to discuss these findings with each other. A hike was organized in the afternoon of the third day.

The seminar was a success: all participants actively participated in the working groups and the discussions, and went home with new ideas and new collaborators. This report gathers the abstracts of the 6 tutorials and the reports of the working groups, which we consider as valuable contributions towards a full-fledged roadmap.

**Seminar** August 21–26, 2022 – <https://www.dagstuhl.de/22342>

**2012 ACM Subject Classification** Artificial Intelligence → Natural Language Processing; Security and Privacy → Human and Societal Aspects of Security and Privacy; Security and Privacy → Software and Application Security; Security and Privacy → Database and storage security

**Keywords and phrases** Privacy, Speech and Language Technology, Privacy Enhancing Technologies, Dagstuhl Seminar

**Digital Object Identifier** 10.4230/DagRep.12.8.60

---

\* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Privacy in Speech and Language Technology, *Dagstuhl Reports*, Vol. 12, Issue 8, pp. 60–102

Editors: Simone Fischer-Hübner, Dietrich Klakow, Peggy Valcke, Emmanuel Vincent



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Executive Summary

*Simone Fischer-Hübner (Karlstad University, SE)*

*Dietrich Klakow (Saarland University – Saarbrücken, DE)*

*Peggy Valcke (KU Leuven, BE)*

*Emmanuel Vincent (Inria – Nancy, FR)*

License © Creative Commons BY 4.0 International license  
© Simone Fischer-Hübner, Dietrich Klakow, Peggy Valcke, Emmanuel Vincent

In the last few years, voice assistants have become the preferred means of interacting with smart devices and services. Chatbots and related language technologies such as machine translation or typing prediction are also widely used. These technologies often rely on cloud-based machine learning systems trained on speech or text data collected from the users. The recording, storage and processing of users' speech or text data raises severe privacy threats. This data contains a wealth of personal information about, e.g., the personality, ethnicity and health state of the user, that may be (mis)used for targeted processing or advertisement. It also includes information about the user identity which could be exploited by an attacker to impersonate him/her. News articles exposing these threats to the general public have made national headlines.

A new generation of privacy-preserving speech and language technologies is needed that ensures user privacy while still providing users with the same benefits and companies with the training data needed to develop these technologies. Recent regulations such as the European General Data Protection Regulation (GDPR), which promotes the principle of privacy-by-design, have further fueled interest. Yet, efforts in this direction have suffered from the lack of collaboration across research communities. This Dagstuhl Seminar was the first event to bring 6 relevant disciplines and communities together: Speech Processing, Natural Language Processing, Privacy Enhancing Technologies, Machine Learning, Human Factors, and Law.

After 6 tutorials given from the perspective of each of these 6 disciplines, the attendees gathered into cross-disciplinary working groups on 6 topics. The first group analyzed the privacy threats and the level of user control for a few case studies. The second group focused on anonymization of unstructured speech data and discussed the legal validity of the success measures developed in the speech processing literature. The third group devoted special interest to vulnerable groups of users in regard to the current laws in various countries. The fifth group tackled the design of privacy attacks against speech and text data. Finally, the sixth group explored the legal interpretation of emerging privacy enhancing technologies.

The reports of these 6 working groups, which are gathered in the following, constitute the major result from the seminar. We consider them as a first step towards a full-fledged interdisciplinary roadmap for the development of private-by-design speech and language technologies addressing societal and industrial needs.

## 2 Table of Contents

### Executive Summary

*Simone Fischer-Hübner, Dietrich Klakow, Peggy Valcke, Emmanuel Vincent . . . . .* 61

### Overview of Talks

Speech privacy  
*Emmanuel Vincent . . . . .* 63

Privacy-enhancing natural language processing  
*Pierre Lison . . . . .* 63

Privacy from a security perspective  
*Meiko Jensen . . . . .* 63

Privacy issues and mechanisms in machine learning  
*Olga Ohrimenko . . . . .* 64

Human factors in privacy  
*Zinaida Benenson . . . . .* 64

Voice and speech: the perspective of legal scholars  
*Lydia Belkadi, Abdullah Elbi, Peggy Valcke, Els Kindt . . . . .* 65

### Working Groups

Case studies and user interaction  
*Zinaida Benenson, Abdullah Elbi, Zekeriya Erkin, Natasha Fernandes, Simone Fischer-Hübner, Ivan Habernal, Els Kindt, Anna Leschanowsky, Pierre Lison, Christina Lohr, Emily Mower Provost, Jo Pierson, David Stevens, Francisco Teixeira, Shomir Wilson . . . . .* 66

Metrics for anonymization of unstructured datasets  
*Lydia Belkadi, Martine De Cock, Natasha Fernandes, Katherine Lee, Christina Lohr, Olga Ohrimenko, Andreas Nautsch, Laurens Sion, Natalia Tomashenko, Marc Tommasi, Peggy Valcke, Emmanuel Vincent . . . . .* 73

Vulnerable groups and legal considerations  
*Lydia Belkadi, Meiko Jensen, Dietrich Klakow, Katherine Lee, Olga Ohrimenko, Jo Pierson, Emmanuel Vincent . . . . .* 80

Privacy attacks  
*Abdullah Elbi, Anna Leschanowsky, Pierre Lison, Andreas Nautsch, Laurens Sion, Marc Tommasi . . . . .* 85

Privacy enhancing technologies  
*Martine De Cock, Zekeriya Erkin, Simone Fischer-Hübner, Meiko Jensen, Dietrich Klakow, Francisco Teixeira . . . . .* 90

Uncertain legal interpretation(s) for emerging PETs  
*Lydia Belkadi, Peggy Valcke . . . . .* 99

**Conclusion . . . . .** 101

**Participants . . . . .** 102

## 3 Overview of Talks

### 3.1 Speech privacy

*Emmanuel Vincent (Inria – Nancy, FR) [emmanuel.vincent@inria.fr](mailto:emmanuel.vincent@inria.fr)*

License  Creative Commons BY 4.0 International license  
© Emmanuel Vincent

Large-scale collection, storage, and processing of speech data poses severe privacy threats. Indeed, speech encapsulates a wealth of personal data (e.g., age and gender, ethnic origin, personality traits, health and socio-economic status, etc.) which can be linked to the speaker's identity via metadata or via automatic speaker recognition. Speech data may also be used for voice spoofing using voice cloning software. In this tutorial, I provide an overview of privacy preservation solutions for speech data, with a focus on voice anonymization. I define the voice anonymization task and evaluation metrics, and outline solutions based on voice conversion and differential privacy. I also briefly mention federated learning, and conclude by stating open questions for future research.

### 3.2 Privacy-enhancing natural language processing

*Pierre Lison (Norsk Regnesentral – Oslo, NO) [plison@nr.no](mailto:plison@nr.no)*

License  Creative Commons BY 4.0 International license  
© Pierre Lison

This tutorial describes the main aspects of privacy-enhancing techniques developed in the field of Natural Language Processing. We first explain the main privacy risks that may arise from processing text or training natural language processing models. We then review a number of privacy-enhancing techniques, in particular text sanitization, text obfuscation, text rewriting and synthesis, and privacy-preserving training of natural language processing models. We also discuss a number of open challenges and research questions.

### 3.3 Privacy from a security perspective


*Meiko Jensen (Karlstad University, SE) [meiko.jensen@kau.se](mailto:meiko.jensen@kau.se)*

License  Creative Commons BY 4.0 International license  
© Meiko Jensen

From a technical or security perspective, privacy has specific connotations and definitions beyond legal or societal dimensions. Especially in the process of designing IT systems, such as AI-based natural language processing systems, these challenges must be addressed appropriately, based on a common understanding of the exact notions of each domain. In this tutorial, I provide some technical definitions of common privacy-related concepts (such as anonymity, or the difference between data and information), and I explain the approach of the protection goals for privacy engineering as an interdisciplinary effort to harmonize privacy considerations at the intersection of law, society, and information technology.

### 3.4 Privacy issues and mechanisms in machine learning

*Olga Ohrimenko (University of Melbourne, AU) oohrimenko@unimelb.edu.au*

License  Creative Commons BY 4.0 International license  
© Olga Ohrimenko

Machine learning models, including those that process text, can leak information about their training data. This has been demonstrated by several attacks (e.g., identifying whether a record in the training dataset, extraction of phrases). Algorithms and mechanisms for protecting training data can be grouped into those that can protect against a data collector and those that protect from a user of the trained model (e.g., for text generation). Secure hardware, cryptographic techniques and local differential privacy can be used for the former setting and have a set of tradeoffs in terms of guarantees, assumptions, and performance. The latter group includes central differential privacy. Though differential privacy is seeing adoption in practice, its applicability for text and speech is an open question and depends on a unit of privacy one is interested in protecting (e.g., a user, a phrase, an utterance, voice) that may be difficult to define.

### 3.5 Human factors in privacy

*Zinaida Benenson (Friedrich-Alexander-Universität – Erlangen, DE) zinaida.benenson@fau.de*

License  Creative Commons BY 4.0 International license  
© Zinaida Benenson

This tutorial discusses how people make decisions about sharing or withholding their data towards commercial organization, governmental organizations and individuals. Unfortunately, we cannot expect people to act in their best interest in this domain. Privacy decisions are subject to many psychological effects: they are heavily dependent on context (who asks, in which order, how the request is framed) and to well-known behavioral biases such as unrealistic optimism and immediate gratification. Moreover, people overestimate risks of terrorism and similar high-emotion threats, which makes them susceptible to the rhetoric of “surveillance for the greater good”, no matter whether this surveillance actually reduces risk. Additionally, ubiquitous presence of IoT devices in private and public spaces raises new issues concerning interpersonal privacy: how to negotiate differing privacy preferences of different users, such as regular inhabitants of smart homes and bystanders.

### 3.6 Voice and speech: the perspective of legal scholars

*Lydia Belkadi (KU Leuven, BE) lydia.belkadi@kuleuven.be*

*Abdullah Elbi (KU Leuven, BE) abdullah.elbi@kuleuven.be*

*Peggy Valcke (KU Leuven, BE) peggy.valcke@kuleuven.be*

*Els Kindt (KU Leuven, BE) els.kindt@kuleuven.be*

**License**  Creative Commons BY 4.0 International license  
© Lydia Belkadi, Abdullah Elbi, Peggy Valcke, Els Kindt

The entanglement of different attributes within speech and text snippets raises important challenges from a legal perspective. For example, it is unclear how speech or text snippets should be defined from a legal perspective or how to apply existing legal definitions. Similarly, this entanglement implies considerable contradictions with data protection principles, such as data minimization and purpose limitation. In other words, snippets may reveal more data than is necessary for a given purpose (e.g., text and typing patterns in language processing). In addition, from a legal perspective, special attention must be given to the concept of vulnerability where the wide spread use of speech technologies may create new type of vulnerabilities. In some situations, the users' right to privacy may conflict with the voice technology company's legal requirements. For example, if the voice technology company collects speech or text data suggesting that a crime (e.g., child abuse) or a life-threatening danger (e.g., heart attack) has taken place, should it report it to the relevant authority, thereby violating the user's privacy? Questions identified during the law tutorial included:

- Are there practices that should be prohibited? What are red lines to the use of voice snippets? (in light of existing/possible safeguards) What are risky applications? (e.g., emotions – what is technically possible or not possible?)
- Can we work towards a common terminology / vocabulary to carry out risk assessments / Data Protection Impact Assessments?
- Should we consider “outliers” (i.e., people whose voice is more identifiable than others) as a new vulnerable group? Novel Speech and Language Technologies as creating new types of vulnerabilities?
- Ethical / moral questions: shall the staff of voice technology companies intervene in the situations when they pick up worrying situations while screening users' voice data? Shall large-scale users' speech and language data be used as legal evidence?

## 4 Working Groups

### 4.1 Case studies and user interaction

Zinaida Benenson (Friedrich-Alexander-Universität – Erlangen, DE) [zinaida.benenson@fau.de](mailto:zinaida.benenson@fau.de)

Abdullah Elbi (KU Leuven, BE) [abdullah.elbi@kuleuven.be](mailto:abdullah.elbi@kuleuven.be)

Zekeriya Erkin (TU Delft, NL) [z.erkin@tudelft.nl](mailto:z.erkin@tudelft.nl)

Natasha Fernandes (Macquarie University – Sydney, AU) [natasha.fernandes@mq.edu.au](mailto:natasha.fernandes@mq.edu.au)

Simone Fischer-Hübner (Karlstad University, SE) [simone.fischer-huebner@kau.se](mailto:simone.fischer-huebner@kau.se)

Ivan Habernal (TU Darmstadt, DE) [ivan.habernal@tu-darmstadt.de](mailto:ivan.habernal@tu-darmstadt.de)

Els Kindt (KU Leuven, BE) [els.kindt@kuleuven.be](mailto:els.kindt@kuleuven.be)

Anna Leschanowsky (Fraunhofer IIS – Erlangen, DE) [anna.leschanowsky@iis-extern.fraunhofer.de](mailto:anna.leschanowsky@iis-extern.fraunhofer.de)

Pierre Lison (Norsk Regnesentral – Oslo, NO) [plison@nr.no](mailto:plison@nr.no)

Christina Lohr (Friedrich-Schiller-Universität – Jena, DE) [christina.lohr@uni-jena.de](mailto:christina.lohr@uni-jena.de)

Emily Mower Provost (University of Michigan – Ann Arbor, US) [emilykmp@umich.edu](mailto:emilykmp@umich.edu)

Jo Pierson (Free University of Brussels, BE) [jo.pierson@vub.be](mailto:jo.pierson@vub.be)

David Stevens (Gegevensbeschermingsautoriteit – Brussels, BE) [david.stevens@apd-gba.be](mailto:david.stevens@apd-gba.be)

Francisco Teixeira (Instituto Superior Técnico – Lisbon, PT) [francisco.s.teixeira@tecnico.ulisboa.pt](mailto:francisco.s.teixeira@tecnico.ulisboa.pt)

Shomir Wilson (Pennsylvania State University – University Park, US) [shomir@psu.edu](mailto:shomir@psu.edu)

**License** © Creative Commons BY 4.0 International license

© Zinaida Benenson, Abdullah Elbi, Zekeriya Erkin, Natasha Fernandes, Simone Fischer-Hübner, Ivan Habernal, Els Kindt, Anna Leschanowsky, Pierre Lison, Christina Lohr, Emily Mower Provost, Jo Pierson, David Stevens, Francisco Teixeira, Shomir Wilson

Two separate working groups were initially created on case studies, stakeholders, risks, and benefits on the one hand, and on user control on the other hand. After the first discussion session, they decided to merge. Hence we present their joint outcomes below.

#### 4.1.1 Existing uses of speech and language technology

Speech and natural language are fundamental to human communication, and they serve as conduits for enormous amounts of personal information. Language technology users share information across a spectrum of levels of privacy sensitivity, from mild to acutely strong.

Uses of speech and language technologies emerged early in the era of digital computers and in recent years they have become ubiquitous. We list some currently existing technologies to motivate the discussion that follows. Many of these may involve a combination of spoken language, acoustics, or written language:

- call center monitoring, e.g., to evaluate the performance of call center agents,
- automated phone menu systems,
- medically-focused technologies, e.g., for diagnosis or tracking symptom severity,
- language learning, e.g., apps for learning to read or speak a second language,
- voice assistants, such as Amazon’s Alexa and Apple’s Siri,
- machine translation between natural languages,
- law enforcement and security, e.g., to detect malicious activity,
- web search, which (like many items in this list) could be text or speech,
- search specific to websites or services, such as on Amazon.com or Facebook,
- large-scale analysis of documents, such as legal documents like court records or laws,
- online social networks, such as Twitter and TikTok,
- writing support services, such as Grammarly.

### 4.1.2 Stakeholders

Stakeholders in speech and voice technology include:

- the individual, i.e., the person whose voice or language are being processed, also referred to as the data subject (in some cases, this individual might actually also be the user of a speech or language technology or only the data subject),
- other individuals, e.g., whose voices are incidentally included in speech audio recordings, or who may be the subject of text written by the individual,
- the first-party service provider, with whom the individual directly interacts,
- third parties (i.e., external to the user and the first party) that the first party shares an individual's data with to fulfill aspects of their service,
- third parties that the first party shares an individual's data with for nonessential purposes, e.g., marketing-focused data brokers,
- government entities, including public agencies and law enforcement,
- the individual's employer or school, if applicable,
- data protection authorities.

This list is not meant to be comprehensive and other stakeholders are likely to exist.

#### 4.1.2.1 Data provenance

We specify three common categories of data sources, acknowledging that there may be more:

- *input data*, that is information disclosed through participation by the individual and provided by the individual to the speech and/or language application,
- *inferred data*, that is data created by the application automatically or manually by labels/annotations of the data received, where the labels/annotations were not obtained by the participation of the individual,
- *metadata*, that is technical information associated with either the input data or inferred data, e.g., time stamps, location data, etc.

*Note:* very recently (August 1st 2022) the Court of Justice of the European Union ruled that the level of protection is the same for sensitive data directly provided by the individual itself, as for other types of (non-sensitive) personal data from which “sensitive information” (e.g., political preference, sexual preference, etc., see Article 9 of the GDPR) can be inferred. Applied to voice technology, this means that the higher standards of protection (as sensitive data, e.g., “explicit consent” vs. “normal consent”) would be applicable to all voice and language technologies.<sup>1</sup>

#### 4.1.2.2 Preliminary categorization

As a next step, we have trimmed down the list of uses of speech and language technology to a more workable number of types of uses from a data protection risk-based perspective. In this respect, two criteria of risk seem particularly relevant. First, we take into account the situations in which the processing will take place (e.g., on-device). This allows us to describe risk in terms of the likelihood of information leakage. The second criterion we applied is the potential combination of data (because combinations of voice related data with other types of personal data are likely to be more problematic from a data protection point of view).

<sup>1</sup> <https://curia.europa.eu/juris/document/document.jsf?text=&docid=263721&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=481514>



Finally, we also consider the number of parties that can have access to the personal data as an indicator of increasing risk to the private sphere of the individual involved.

Applying these criteria, we identify the following three categories of situations in which speech and language *data can be processed*:

1. locally on a user device, also referred to as “on-device” processing, where input data and inferred data (see definition of terms) does not leave the device (maximum user control and most limited number of parties involved),
2. networked or connected services in which input data and/or inferred data are transmitted from the device that recorded input data (e.g., provided by a commercial service provider, for example online communication between users),
3. processing of data without active intervention or request of the individual (e.g., in the public domain by a public authority, for example usage of voice enabled cameras in public areas, or using voice technology in employer-employee context).

We are fully aware that our proposed categorization has limits. First, it presupposes the availability of a significant amount of information about the technical set-up of a product or a service. Such information might not always be easily or publicly accessible. Second, it is not unlikely that a particular speech or voice product or service might fall in more than one category (example 1: checking medical conditions might be done by a combination of processing locally on a device, while also processing some part of the data in a networked mode; example 2: the processing of wake-up commands by Alexa, both in a local and networked mode).

We identify physical scopes of *data storage*: on a local device (typically one the user interacts with directly) or on remote servers (including but not limited to cloud storage). A separate dimension is the intended scope of access, which may include an arbitrary subset of these options: the user only, the service provider, third parties that the user specifically designates, and the general public.

The case scenarios implementing speech and language technology are numerous. For the purposes of the discussion below, we identified three specific examples, which could stand for three different categories of use cases, based on factors such as user control, parties involved in the processing activities and power and information asymmetry:

*Scenario. 1: Speech diagnosis by health practitioners:* In a doctor–patient relationship, speech and language technology can be used to aid in the diagnosis of particular disorders, determination of treatment and/or monitoring any progress of medication and treatment.

*Scenario. 2: Online language learning service:* A mobile application (“App”) that provides a user with a curriculum to learn to write and/or speak a new language.

*Scenario. 3: Recording of voice and speech in public places:* In the last decades, cameras have emerged in public areas. Recently, some cities are experimenting with the additional registration of audio by these devices in order to fight noise pollution<sup>2</sup> or for public safety or policing purposes (e.g., recognition of aggression in public spaces)<sup>3</sup>. The usage of voice enabled cameras in public contexts is a case study of particular concern.

In addition, we also discuss some specific needs of scientific research in the public interest, in particular the need for available data (both personal and non-personal data) such as for training speech and language models.<sup>4</sup> Societies have become data economies with increasing

<sup>2</sup> <https://www.vrt.be/vrtnws/nl/2021/09/24/genk/>

<sup>3</sup> <https://www.ed.nl/eindhoven/netwerk-van-hypermoderne-camera-s-op-stratumseind-in-eindhoven-gaat-politie-helpen-a1e8acee/?referrer=https%3A%2F%2Fwww.google.com%2F>

<sup>4</sup> See e.g., EU Commission, A European Strategy for data, COM/2020/66 final, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066>.

needs for data, for the benefit of people, organizations, economy and society progress as a whole. Specific safeguards however are needed and are moreover legally required under the European data protection legislation to protect information about identified and identifiable individuals. The usual safeguards of anonymization and pseudonymization are relevant and briefly discussed hereunder, but also the limitation thereto.

### 4.1.3 User control and privacy threats

User control is at the core of data protection. Individuals shall be given the choice as for the collection of additional information and any consent shall be in a granular way.<sup>5</sup>

While individuals are given the option to agree (opt-in) with the collection and use of additional information extraction from the speech and language application, there is a profound risk that their choice will not be taken into account, because

- the algorithmic learning models may already have information about demographics, etc.,
- the company or entity uses different labels/annotations.

The latter issue may lead the company or entity to avoid or not acknowledging that specific inferred information is processed. This may seem problematic, but in the end, it will however remain the responsibility of the company/entity to label the inferred information correctly and to respect the choice of the individual. The first issue, however, remains problematic, especially in an increasingly “connected world” with dominant players. Cross-correlation of data from different platforms requires unambiguous consent.

Additionally, users might not be able to make informed choices due to misleading phrasing and confusing interfaces fraught with dark patterns, which is already happening on large scale with cookie consent notices [1]. The companies will be tempted to use dark patterns and nudging towards privacy-decreasing choices also in case of consent notices for language and speech processing, as their business models depend on this data, just like in the case of cookies.

At the same time, user control may not be sufficient in case of privacy interferences, when applications are invading in the “private sphere”, such as in use case 3. Individuals are entitled to respect privacy even in public places, and even if they would be public persons. At the same time, “privacy is a broad term, not susceptible of a definition”. It encompasses a wide array of interests, including the right to personal development and to engage in relationships, to meet and to engage with other people. Individuals also have (some degree of) privacy when conducting professional activities and are entitled to protect their identity. And – also very importantly – privacy may be needed to exercise fundamental rights, including the right to free speech or to protest. Privacy is therefore inherently linked with freedom.

Any risk of applications limiting privacy shall therefore be assessed at the design phase of each and any voice, speech and text application. The concerns shall be addressed hence before development, right from the start and, for example, by using PETs or organizational measures (“privacy and data protection by design”). If this would not be sufficient, only limited exceptions to the fundamental right to privacy are possible but only in as far as necessary (“is it the last measure that can be effective, e.g., to curb public threat”) and proportionate (“is it in proportion with the legitimate goal to be reached?”) in democratic societies, and a sufficiently precise law is adopted to allow the interference.

<sup>5</sup> See Article 29 Working Party, Opinion on consent, [https://en.wikipedia.org/wiki/Article\\_29\\_Data\\_Protection\\_Working\\_Party](https://en.wikipedia.org/wiki/Article_29_Data_Protection_Working_Party).

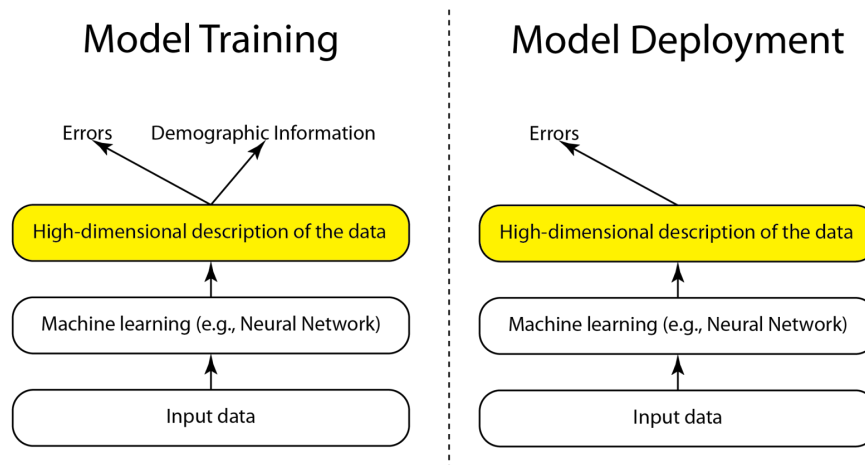
#### 4.1.3.1 User privacy in speech and language technology

We draw a distinction between input data and inferred data (see above). Inferences may include characteristics of an individual that can be automatically extracted from their input data, including, but not limited to, culture, race, age, gender identity, socioeconomic status, education, marital or parental status, health information, location, emotion, and stress. Inferred data does not have to be human interpretable. A more detailed discussion on this can be found in the section on PETs.

One way for a computing system to gather information about a user is to ask them directly. In that case, the terms of use guide how these characteristics are used and shared. However, when the input data include audio, speech, and text, and these characteristics are *inferred* rather than *disclosed*, it may become less clear how or if the inferred characteristics, the inferred data, can be reused.

One path to protect the consumer’s non-disclosed information is to place protections around the inference of the characteristics, for example noting that emotion or gender identity should not be inferred. This is in line with the concept of sticky policies and privacy rights management, defined as “a form of digital rights management involving licenses to personal data”. These policies describe what can and cannot be done with a given data resource. However, due to the complexity of machine learning algorithms, it is difficult to enforce this.

For example, consider an application designed to teach a user to speak a foreign language. It may be advantageous to understand how gender identity, culture, age, or many other demographic factors influence the types of errors that may be observed. Therefore, the company may be incentivized to train algorithms that learn to recognize errors (e.g., mistakes made in pronunciation, grammar, or word choice) and how those errors overlap with these demographic identifiers. To do so they would collect data that includes both errors and demographic identifiers and train a system to jointly predict both errors and the demographic identifiers (see Fig. 1, left). This would result in a predictive model and a high-dimensional description of the data (see the yellow box in Fig. 1).



■ **Figure 1** Model training and deployment.

When the model is deployed (see Fig. 1, right), in line with the consumer protections, it would not include the prediction of the demographic characteristics. Thus, it would not be inferring demographic characteristics because the demographic information classifier is not included. However, the same yellow embedding, the embedding that distills out the

demographic characteristics, would be generated when the model was deployed (note: this is true even when demographic information is not included as a classification target). As such, demographic characteristics would be included in the learned numeric representation of the data. These representations could then be automatically clustered (grouped) to identify similar users. Thus, although the exact information about their demographic characteristics is not known, inferences about these characteristics will be.

These inferred data have value. They can be aggregated across data sources to form detailed user profiles that may guide decision making ranging from advertising (which products should be displayed to which users, when?), insurance (who is at risk of serious, and expensive, illness?), mortgage loans (who is higher or lower risk), job hiring (who has characteristics that a company may find (un)desirable), law enforcement, and more. The question is then, what, if anything, should be done to control how these inferences are reused?

We highlight this challenge in Fig.2 using the example of a language learning app, one that takes in acoustic information and provides feedback to a user to promote the user's language mastery. We assume that the app requires audio information and the ability to extract speech-language information (note the red exclamation point in the matrix). The company would like to retain this information to improve the model's performance and the app's behavior. The company would also like to use this information to build a user profile, a mechanism that would allow the user to automatically advance through the app, given mastery. The company may desire text feedback, although this is not required. However, there are no mechanisms in place that safeguard the inference of the user's characteristics either within the functionality of the system itself or outside of the company, or organization, that has collected this information. We highlight this challenge in the matrix, using a box that notes "application of privacy regulations is unclear". We borrow inspiration for this matrix from prior work on consumer privacy nutrition labels [2].

#### 4.1.3.2 User awareness and concerns about inferred information

As outlined above, highly sensitive information can be inferred from speech and language data: age, gender, ethnicity, geographical origin, emotional states, physical states (e.g., intoxication level), health-related information, intention to deceive [4]. Respective privacy threats can be roughly divided into impersonation and profiling. Impersonation refers to spoofing user identity, e.g., for authentication purposes, but also for spreading fake news and defamation. Profiling facilitates targeted advertising (including political marketing), but also discrimination, e.g., in language-based services such as call centers, or in job application processes. Additional privacy threats arise from language models for text and speech processing, as neural network language models can memorize the training data and reveal secrets from it. See more information in the section on possible attacks.

In user studies on privacy in smart homes, users generally express concerns about storage of their voice recordings by providers. For example, Malkin et al. [5] showed that unlimited storage of voice recordings, which is the default option for Amazon Alexa and Google Home, does not match well with users' expectation that this data should only be stored for short periods, and then deleted. At the same time, voice data was not considered to be particularly sensitive, and over 70% of participants reported that they have never had privacy concerns about their devices.

Yet, the general public seems to be poorly informed about possible inferences from text and voice processing and threats originating from these. To the best of our knowledge, Kröger et al. [3] were the first to explicitly investigate user awareness of and concerns about inferences from voice recordings. They asked a representative sample of the UK population

An Example for a Language Learning App: learn to speak a foreign language									
Input Data	Inferred Data	Within Organization (can be very broad)				Outside of Organization			
		Strictly fulfilling the service	Research and development (algorithm improvements)	Profiling	Marketing	Marketing	Profiling in Aggregation	Other categories	Public forums
Audio	—	!	Choice	Choice	-	Likely to be considered as reuse under existing regulations			
Speech language	—	!	Choice	Choice	-				
Text	—	-	-	-	-				
—	Socioeconomic Status	These are thought to be individual choices, to which a user can either opt-in or opt-out.				Application of privacy regulations is unclear			
—	Health Information								
—	Age								
—	Gender Identity	The reality is that we have very little control over these decisions because of complex machine learning solutions that have already learned these correlations.							
—	Native Language								
—	Accent								
—	Location								
—	Emotion								
—	Stress								

<b>Key</b>	
!	We will use your information in this way
-	Not Used: we will not collect or use your information in this way
Choice	User choice: 1) we will not use your information in this way unless you opt-in OR we will use your information in this way unless you opt-out

■ **Figure 2** Example language learning app.

(n=683) to indicate how aware they are of three types of inferences: demographic data (age, gender, geographic origin), short- and medium-term states (e.g., intoxication, sleepiness, moods and emotions) as well as personal traits (mental and physical health, personality traits). Overall awareness level was quite low and depended on the inference type. Whereas awareness of the demographic inferences was the highest (almost 50% of respondents reported to be at least somewhat aware of it), only around 20% of respondents reported at least some awareness of the personal trait's inferences, with the awareness of short- und medium-term states inferences being in-between. Concern level about inferences was mixed, with around 40% of participants reporting to be concerned, and approximately the same percentage reporting to be unconcerned. When asked to justify their concern level, participants provided free-text answers that indicated, e.g., well-known privacy misconceptions such as "I've got nothing to hide" [6], a lack of knowledge about possible misuse of inferred data, but also the perception that benefits of voice-based technologies outweigh their dangers.

#### 4.1.3.3 Moving forward

User awareness and control are very complex and subject to well-known behavioral biases. For example, Acquisti et al. [7] showed in a series of experiments that users can be manipulated towards greater information disclosure by distractions such as small delays. Furthermore, they showed that increased perceived control over the release of information also increases risky behavior, leading to higher information disclosure. As a result, awareness may have

only limited (or even adverse!) impact on safeguarding users' speech and language data. Yet, users must receive this information in a manner that is comprehensible and devoid of nudging and dark patterns. They should be able to know what happens with the data and what can be inferred. Further, regulating bodies should be made aware, or increasingly aware, of the complexities in this space. However, users' and policy makers' awareness alone will not solve the problem. We must identify additional regulations around the reuse of inferred data when these data contain personally identifiable information or otherwise personal data.

## References

- 1 Lorrie Faith Cranor. Cookie monster. *Communications of the ACM*, 65(7):30–32, 2022.
- 2 Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. A “nutrition label” for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–9, 2009.
- 3 Jacob Leon Kröger, Leon Gellrich, Sebastian Pape, Saba Rebecca Brause, and Stefan Ullrich. Personal information inference from voice recordings: User awareness and privacy concerns. *Proceedings on Privacy Enhancing Technologies*, (1):6–27, 2022.
- 4 Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. Privacy implications of voice and speech analysis—information disclosure by inference. In *IFIP International Summer School on Privacy and Identity Management*, pages 242–258. Springer, Cham, 2019.
- 5 Nathan Malkin, Joe Deatruck, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies*, (4):250–271, 2019.
- 6 Daniel J. Solove. I've got nothing to hide and other misunderstandings of privacy. *San Diego L. Rev.*, 44:745, 2007.
- 7 Alessandro Acquisti, Idris Adjerid, and Laura Brandimarte. Gone in 15 seconds: The limits of privacy transparency and control. *IEEE Security & Privacy*, 11(4):72–74, 2013.

## 4.2 Metrics for anonymization of unstructured datasets

Lydia Belkadi (KU Leuven, BE) [lydia.belkadi@kuleuven.be](mailto:lydia.belkadi@kuleuven.be)

Martine De Cock (University of Washington – Tacoma, US) [mdecock@uw.edu](mailto:mdecock@uw.edu)

Natasha Fernandes (Macquarie University – Sydney, AU) [natasha.fernandes@mq.edu.au](mailto:natasha.fernandes@mq.edu.au)

Katherine Lee (Google Brain & Cornell University – Ithaca, US) [katherinelee@google.com](mailto:katherinelee@google.com)

Christina Lohr (Friedrich-Schiller-Universität – Jena, DE) [christina.lohr@uni-jena.de](mailto:christina.lohr@uni-jena.de)

Andreas Nautsch (Avignon Université, FR) [andreas.nautsch@univ-avignon.fr](mailto:andreas.nautsch@univ-avignon.fr)

Laurens Sion (KU Leuven, BE) [laurens.sion@kuleuven.be](mailto:laurens.sion@kuleuven.be)

Natalia Tomashenko (Avignon Université, FR) [natalia.tomashenko@univ-avignon.fr](mailto:natalia.tomashenko@univ-avignon.fr)

Marc Tommasi (University of Lille, FR) [marc.tommasi@univ-lille.fr](mailto:marc.tommasi@univ-lille.fr)

Peggy Valcke (KU Leuven, BE) [peggy.valcke@kuleuven.be](mailto:peggy.valcke@kuleuven.be)

Emmanuel Vincent (Inria – Nancy, FR) [emmanuel.vincent@inria.fr](mailto:emmanuel.vincent@inria.fr)

License © Creative Commons BY 4.0 International license

© Lydia Belkadi, Martine De Cock, Natasha Fernandes, Katherine Lee, Christina Lohr, Olga Ohrimenko, Andreas Nautsch, Laurens Sion, Natalia Tomashenko, Marc Tommasi, Peggy Valcke, Emmanuel Vincent

### 4.2.1 Introduction

Article 32 of the GDPR requires data controllers and processors to implement “appropriate technical and organizational measures to ensure a level of security appropriate to the risk”. Such measures may include pseudonymization, encryption, the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and



services. Reference is also made to processes for “regularly testing, assessing and evaluating the effectiveness of technical and organizational measures for ensuring the security of the processing”.

Notions like confidentiality and integrity have been borrowed from the fields of security and privacy engineering, and their meaning is often hard to grasp for lawyers when implementing the law or assessing compliance of systems and applications. This underlines the need to bridge technical and legal vocabularies and methods, something which is also particularly important in the context of Article 25 of the GDPR (data-protection-by-design).

To facilitate this “translation”, Data Protection Authorities have issued guidance documents specifying what data protection entails: which privacy and security goals does it intend to achieve, and what do these mean in terms of risks and metrics? One example of this are the six protection goals developed by the German Data Protection Authority or the Article 29 Working Party Opinion 05/2014 on Anonymisation Techniques.<sup>6</sup>

In the latter, the Article 29 Working Party (the predecessor of the European Data Protection Board) offers guidance on key notions on the context of pseudonymization and anonymization techniques, to assess whether they can provide appropriate privacy guarantees. It is, however, clear from the use of “database” throughout the document that Opinion 05/2014 has been written with structured data in mind. Given that text and speech are mainly unstructured data, the question arises to what extent the parameters used in existing risk assessment frameworks are still appropriate. What do notions like *singling out* or *linkability* mean in a speech and text context? Are they still relevant to capture and measure risks to privacy? What are the shortcomings? What are possible alternative notions?

The question is particularly relevant given that the GDPR applies only to the processing of personal data, defined as “any information relating to an identified or identifiable natural person (“data subject”); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person”. By contrast, Recital 26 of the GDPR considers anonymous information as “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”. To 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.

#### 4.2.2 Structured vs. unstructured data

To guide the discussion below, we first explain the difference between structured and unstructured data. Structured data is typically stored in databases where each row of the database contains the data of an individual, and such data is structured according to named

---

<sup>6</sup> [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216\\_en.pdf](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf)

attributes (columns) which have a clear (identified) meaning. Unstructured data, including text, speech and images, is stored in such a way that attributes of the data are not explicitly identified aside from (possibly structured) metadata associated with the data snippet.

The language of data protection appears to assume that data is stored in a structured format, such that individuals can be identified with attributes which are explicitly stored in the data. Words associated with structured data include “records”, “attributes” and “databases”. Moreover, structured data assumes the concept of a data subject or individual, which corresponds to the person identified by a row of the database.

We think the reason why these notions become increasingly complex is related to the fact that there are several layers in speech/text that need legal protection: 1° the content of the speech/text (which can contain personal data that isn’t necessarily exclusive to the author/speaker); 2° the identity of the author/speaker (derived from the physiological and/or behavioral characteristics of the voice or writing style); 3° other characteristics that you can derive from the voice/writing style (like gender, mental state, etc.).

In unstructured data such as speech and text, the notion of an individual or data subject is ambiguous, as it may refer to the individual who produced the data, the individuals mentioned in the data or even other individuals whose identity can be inferred through the data. In addition, attributes are not explicitly recorded but are implicitly embedded in the data. Extracting these attributes is itself an ongoing research task. Finally, the attributes used to link individuals may not be ones that can be explicitly described in human terms; for example, a machine learning system may use attributes of speech that are not necessarily explainable to a person to infer the identity of an individual.

#### 4.2.2.1 Singling out, linkability and inference

The traditional meaning of singling out, linkability and inference has been described by the Article 29 Working Party in its Opinion 05/2014 on Anonymisation Techniques:

- *singling out* corresponds to the “possibility to isolate some or all records which identify an individual in the dataset”;
- *linkability* is the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases). If an attacker can establish (e.g., by means of correlation analysis) that two records are assigned to a same group of individuals but cannot single out individuals in this group, the technique provides resistance against “singling out” but not against linkability;
- *inference* is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

These definitions leave some room for interpretation, even in the case of structured data. For instance, the notion of “group of data subjects” in the definition of linkability should be restricted to groups that are small enough to “nearly” identify the subject or meaningfully make inferences about the subject as opposed to groups such as “male” or “above 25” which are huge. Most forms of data processing rely on classification/grouping hence maintaining the utility of the anonymized data requires that some grouping is allowed according to the intended usage. Similarly, reasonable inferences about anonymized data should be allowed such that the data remains useful for the intended usage.

In addition, the term “inference” has various meanings in different communities; for example, in natural language processing it typically refers to the process of logical deduction, close to the meaning implied in the above. In the statistics community, inference refers to conclusions reached based on probabilistic estimates, different from the meaning above. In



the context of this document (and particular to unstructured data) it may be the case that inferences (in the statistical sense) are required in order to draw inferences (in the meaning implied above). This confusion of terminology should be clarified.

Since speech and text data are unstructured with entangled information, the processes of singling out, linking, and inference emerge as part of one overarching process that cannot be clearly distinguished into the three processes that are separable for structured data. If only unstructured data is given, to link attributes, one needs to first infer them (to know which values should be linked); to single out one needs to use inference (to know what to filter by); thus, to single out, one needs to do the opposite of linking, thus linking methodology is implied (discarding a hypothetical association/linkage requires to assess the strength of that relation/linkability). For structured data (e.g., processed outcomes only) and mixed structured and unstructured data (e.g., annotated utterances), singling out, linking, and inference can be however different processes. It is only when one starts with nothing but unstructured data, that the three processes become part of one larger coherent process.<sup>7</sup>

Let's try to translate these notions to the text and speech context:

- *Singling out*: Here “individual” could refer either to the creator of the data (the speaker, the author, etc.) or to subjects explicitly referred to in the data (eg. the person who was photographed), or subjects implicitly mentioned in the data (e.g., from background noise). The data protections should apply to all.
  - Speech and text data make it possible to interrogate the data with different questions such that the filtering process of structured data becomes inherently easy using derived data from that speech and text data.
  - It is unclear at this stage how this concept can be quantified. The notion of Predicate Singling Out [1] defined for structured data is not widely agreed-upon yet, it is limited to datasets with a single recording per individual, and its extension to unstructured data raises the question of assessing singling out based on relevant attributes of the data subject as opposed to irrelevant attributes of the data (e.g., the value of the N-th sample in an audio recording) that allow records to be singled-out in the Predicate Singling Out sense but do not identify the data subject in any way.
- *Linkability*: This is the capacity to interrelate (unstructured) data sources, (unstructured) models trained on those data sources, derived (structured) information/attributes, and (even spurious) prior connections to further develop structured records concerning (groups of) individuals.
  - There is further complexity in the speech and language context due to the various interpretations of the “data subject”, who could be the speaker/author, or the subject spoken about, or (in speech) a subject identifiable in the background. Any of these “data subjects” could be linkable to other data records in a data collection. Moreover, background acoustics allow for the characterization of the recording environment.
  - In the framework of the VoicePrivacy challenge<sup>8</sup>, different empirical metrics were compared and assessed: the zero-evidence biometrics recognition assessment “ZEBRA” framework assesses expected and worst-case privacy disclosure motivated from information theory, forensic sciences, and secure communication (cryptography); the unlinkability metric targets the local and global divergence between two protected

<sup>7</sup> Note for machine learning experts: when structured data is involved, inference implies a deduction of unknown elements by applying patterns known from another source; this relates, e.g., to super-interpolation/reconstruction tasks and to joint-intersecting databases.

<sup>8</sup> <https://www.voiceprivacychallenge.org/>

biometric reference datasets when being compared to one another (designed to apply to the ISO/IEC 24745 standard on biometric information protection), and the equal-error rate (easiest explainable metric; explicitly deprecated in ISO/IEC 19795-1:2021 for performance reporting).

- *Inference* is the possibility to deduce, with even weak probability, the (not necessarily correct) value of a (structured) attribute from the values of a set of other attributes of unstructured data (speech, text, ...).
  - Here attributes are derived (implicit) from the data rather than described explicitly as in the case of structured data. It is already known that some speech and text attributes can be measured with a certain level of accuracy. Surprisingly, no attempt has yet been made to assess how well various combinations of such existing attributes identify individuals. In addition, identifying these implicit attributes in the data is still an ongoing research question. This makes it difficult to quantify protections against such inferences in the future.
  - Note that adversaries might collect lots of weakly deduced attributes under aggregated strength to shortlist automatically, and proceed then manually to single-out eventually.

The metrics mentioned in the above discussion are empirical metrics, which rely on attack models and depend on the property of the evaluation datasets. So, risk evaluation depends on the power of these attack models. It therefore depends on continuous assessment to ensure that technology progress in producing attacks is continuously monitored and risks mitigated.

The above discussion also highlights that several metrics have been proposed for singling out and linkability and that they are different, i.e., they do not always agree. While this did make sense for structured data, this does not make sense in the situation when only unstructured data is given. In such a situation, since singling out, linking, and inference are parts of a single overarching process, a single overarching metric would be desirable. Since the adversary remains unknowable, three metrics for structured data (singling-out; linking; inference) and one for unstructured data are relevant to performance reporting; for mixed data, an attacker might use structured data only to validate information gathered from unstructured data, without any attempt of linking structured data explicitly.

As an alternative to the above empirical metrics, formal privacy guarantees may also be considered for data protection impact assessment. The law does not impose one method or the other. Instead, the data controller proposes the method and the Data Protection Authority has responsibility to verify if the assessment is sufficient. Most existing formal guarantees (e.g., differential privacy, k-anonymity, etc.) were also designed with structured data in mind. Hence it is still an open question how to apply formal methods to privacy assessment in unstructured data such as speech and language.

### 4.2.3 Privacy disclosure: risk assessment

It is generally admitted that a Data Protection Impact Assessment should include a determination of the likelihood of a privacy breach (e.g., based on the probability of success) in addition to an impact assessment (made by human judgment) which determines the severity of impact for the data subjects concerned and the society at large. The impact cannot easily be quantified, hence technical means of assessing risk are expected to focus on quantifying the likelihood of the privacy breach. In the case of anonymization, this is the likelihood that the data subject (or a meaningful group of subjects) is re-identified according to the criteria above.

We recommend that the *likelihood* of a privacy breach should be assessed from both the worst-case and average-case perspectives. The likelihood varies depending on the data subject and the actual data. For instance, in the case of speech, some individuals may be

more easily re-identifiable than others, depending on their voice and on the spoken contents [2]. Assessing the worst case means quantifying the maximum breach probability across all subjects and all expected contents, while assessing the average case means quantifying the average breach probability across all subjects and all expected contents. These assessments may have to be done empirically rather than through a formal analysis.

This likelihood analysis should be combined with an impact assessment to determine an appropriate course of action. For example, if the severity of a breach is considered high, then data subjects with a high to moderate likelihood of breach should be removed or appropriate protections put in place to mitigate their risks. In the case of low impact breaches, it may be deemed appropriate to allow individuals with higher likelihood of risk to remain in the data collection. This judgment is expected to be made in coordination with the Data Protection Authority.

While existing methods that are widely used provide a strong support to landscape, navigate, and steer within data protection – for the new world it is to the most –, they are not mandatory by law. The GDPR framework allows for the emergence of new technologies and methodologies that aid better privacy and risk assessment. It should go without saying yet might benefit to be voiced: summaries of the status quo give an impression of how far a community got; one can go beyond.

#### 4.2.4 Renewability revisited. Continuous countermeasure upgrading.

In the biometric information protection standard, “renewability” is defined via revocability: “revocation is required to prevent the attacker from future (or continued) unauthorized access”. There, renewability is used as a security requirement. The goal is to adopt, e.g., a new cryptographic measure without needing to recapture/reacquire data. Here, we extend this notion to information protection for speech and text.

The state-of-the-art in attacking systems continually improves, and we must continue to adopt new countermeasures to unstructured data to continue to offer appropriate protections. The core idea behind the concept word “renewability” is to adopt such new measures to keep up with adversaries getting stronger over time:

- With structured data, the so-far view is to use a better encryption algorithm, enlarge the key size, etc. without needing to reveal or to recapture the unprotected data.
- For unstructured data, especially for speech, no recapture is exactly alike: repeated utterances vary, e.g., a slight change in timbre and background noise changing.
- Regarding upgrading a countermeasure for unstructured data that manipulates the data: data has had been manipulated before. An update that relies on transformation should ensure that there is no inverse function of that transform (such an inverse makes it possible to remove the update); or that there is a considerable amount of effort necessary to obtain a functional inverse. The update process should not be reversible.
- If unstructured data is not stored – processed only –, countermeasures can be uncomplicated in comparison to full data anonymization, e.g., if on-premise computing and access control are sufficient as parts of Data Protection Impact Assessment.
- Continuous upgrading suggests a regularity; not eternal waiting. Different events can be indicators to investigate upgrade strategies: a regular testing and evaluating of Article 32 of the GDPR (timespan to be defined in the Data Protection Impact Assessment); publication of related exploits; any related auditing taking place.
- There needs to be a continuing conversation with Data Protection Authorities for high-risk situations.

- Continuous assessments of the privacy risk of analysis on unstructured data is critical because we do not have formal guarantees for assessment.
- Formal methods always require a set of assumptions and scope. We can grow and improve the scope of formal protections, however, for any data as unstructured as speech and language, continuous assessment of privacy risk is highly recommended.

#### 4.2.5 Defining the “accuracy” of algorithms

Terminology regarding AI is expected to be defined in the Proposal for the Regulation of Artificial Intelligence systems (“AI Act”).<sup>9</sup> To aid the ability of communicating here, we describe contexts towards “accuracy”. Notably, the harmonization of biometric systems as of ISO/IEC 2382-37:2017 does not define “accuracy” whereas a plethora of other metrics is defined to harmonize across the biometric standardization projects. Expert discussions in this ISO/IEC vocabulary harmonization group reached the conclusion that accuracy is not definable – other efforts were reported which attempted the definition of accuracy for over a decade, without success. In our discussions at Dagstuhl, we arrived at the perspective that legal communities (as well as natural language/text processing) operate via formal methods of philosophical reasoning using math, the classical deductive method. When following statistical paradigms, regardless of their type of data (nominal, categorical, etc.), these toolsets are less available, since uncertainty is unavoidable. One can through AI reduce uncertainty regarding associating a specific value to an attribute, but only so much so. To express remaining uncertainty is what “accuracy” implies. The ways of quantifying remaining uncertainty are shaped by the extent to which the entire process chain is covered from sensor capture to policy and governance. AI experts stop at performance reporting, leaving policy thereafter prone. When bringing the disciplines together, “accuracy” could also be reflective about how much of the remaining uncertainty disrupts policy makers to take decisions, or conversely how much information is explainable to decision makers and beneficial for their task at hand.

#### 4.2.6 Assumptions: to promote appropriate uses of privacy methods, state assumptions clearly

Privacy methods all make assumptions about how models will be trained and deployed and how adversaries will act. They are only exhaustive to the extent as their underlying model is fully reflective of the world: countermeasures that have only gone so far but not to the extent that an adversary could go, (formal/logical) loopholes in countermeasure design are readily exploitable. However, to ensure that protections are maximized, privacy method developers must clearly state their assumptions to enable dataset creators and model developers to appropriately meet assumptions or understand the risks of relaxations of the assumptions.

#### References

- 1 Aloni Cohen and Kobbi Nissim. Towards formalizing the GDPR’s notion of singling out. *Proceedings of the National Academy of Sciences*, 117(15):8344–8352, 2020.
- 2 George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Technical report, DTIC Document, 1998.

<sup>9</sup> <https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A52021PC0206>

### 4.3 Vulnerable groups and legal considerations

*Lydia Belkadi (KU Leuven, BE) [lydia.belkadi@kuleuven.be](mailto:lydia.belkadi@kuleuven.be)*

*Meiko Jensen (Karlstad University, SE) [meiko.jensen@kau.se](mailto:meiko.jensen@kau.se)*


*Dietrich Klakow (Saarland University – Saarbrücken, DE) [dietrich.klakow@lsv.uni-saarland.de](mailto:dietrich.klakow@lsv.uni-saarland.de)*

*Katherine Lee (Google Brain & Cornell University – Ithaca, US) [katherinelee@google.com](mailto:katherinelee@google.com)*

*Olga Ohrimenko (University of Melbourne, AU) [oohrimenko@unimelb.edu.au](mailto:oohrimenko@unimelb.edu.au)*

*Jo Pierson (Free University of Brussels, BE) [jo.pierson@vub.be](mailto:jo.pierson@vub.be)*

*Emmanuel Vincent (Inria – Nancy, FR) [emmanuel.vincent@inria.fr](mailto:emmanuel.vincent@inria.fr)*

License  Creative Commons BY 4.0 International license

© Lydia Belkadi, Meiko Jensen, Dietrich Klakow, Katherine Lee, Olga Ohrimenko, Jo Pierson, Emmanuel Vincent

#### 4.3.1 Biometric systems

From a legal perspective, biometric verification (that is verifying the identity of a speaker) systems are often deemed to be not as risky as biometric identification systems (who out of a larger set of known speakers is speaking). Under data protection laws, legal scholars have discussed the definition and legal nature of biometric data. Indeed, Articles 4(14) of the GDPR and 3(13) of the Law Enforcement Directive define biometric data as “personal data resulting from specific technical processing [...] which allow or confirm the unique identification” of an individual. In particular, the definition seems to directly refer to biometric identification (i.e., “allow” the unique identification) and verification (i.e., “confirm” the unique identification) [1]. Article 9 of the GDPR further specifies that only biometric data “processed for the purpose of uniquely identifying” an individual are considered sensitive. In other words, the GDPR does not consider all processing of biometric data as sensitive and excludes verification purposes [2].

This distinction between identification and verification further permeates the risk assessment performed by the European Commission under the AI Act. This regulation aims to set out rules for the development, marketing, and use of AI systems. It further aims to steer AI uptake to reach a high level of protection of public interests (e.g., health, safety, fundamental rights). The AI Act relies on a risk-based framework spanning from unacceptable to minimal risks to support this approach. Accordingly, AI practices entailing severe risks to public interests are prohibited or more strictly regulated.

The current draft of the AI Act considers that biometric verification always entails “minimal risks”, except in the context of migration, asylum and border control management. In particular, AI systems used to verify the authenticity of travel documents and check their security features are considered high-risk (Annex III). This exclusion means that providers, users, and other third parties involved in the supply chain would, in principle, not be subjected to the obligations set out in Articles 16 to 29 (e.g., taking corrective actions in case of non-conformity, information and cooperation with national competent authorities, etc.).

Furthermore, only high-risk AI practices are required to comply with a set of requirements related to the establishment of a risk management system, data governance, technical documentation, record-keeping, transparency and provision of information to users, human oversight and accuracy, robustness and cybersecurity (Articles 9 to 15). These requirements would be applicable to biometric verification systems only on a voluntary basis, through the adoption of codes of conduct (Article 69).

From a technical perspective, linking the risks of biometric verification only to the number of individuals enrolled in a database is criticizable. Indeed, risks still arise even when the database contains a single individual. First, storing biometric identifiers in the cloud as

opposed to the user's device implies that they may be more easily stolen, or that the user might be identified in a situation when they don't want to. Second, the "vocal signature" has been shown to contain a lot more information than biometric identity, which might be inferred [3]. The same risk arises with, e.g., typing patterns associated with text. Third, the boundary between verification and identification is not always clear, e.g., when a smart speaker is used by 5 members of a family, running speaker verification against 5 "vocal signatures" could qualify as a form of identification. The risk should therefore be quantified depending on the usage context, the location where the identifiers are stored, and whether the user is willing to be identified.

### 4.3.2 Beyond identity

Speech and text snippets are complex sources of information conveying more than (biometric) identity. For example, they may reveal speakers' emotional states or health conditions. It is not always possible to dissociate and isolate different attributes captured from individuals, entailing the collection of a wide scope of sensitive personal data. Over time, such collections may also enable the constitution of extensive (e.g., personality) profiles.

Many technical and legal distinctions may be drawn to determine the sensitivity of the collection and processing of speech and text. For example, the collection of a single instance or aggregates of emotional states would have different impacts on concerned individuals. Similarly, the use of aggregates of speech and text snippets for profiling would have distinct risks and benefits depending on the context (e.g., commercial or medical uses). Accordingly, a blanket prohibition of the extraction of specific attributes of speech and text snippets may not be desirable.

At the same time, the entanglement of different attributes within snippets raises important challenges from a legal perspective. For example, it is unclear how speech or text snippets should be defined from a legal perspective or how to apply existing legal definitions. This difficulty was well illustrated in recent legislative debates over the legal concept of "biometric data" under the upcoming AI Act. In particular, the European Parliament is discussing the opportunity to distinguish the concept of "biometric data" and "biometric-based data" to account for processing beyond biometric recognition (e.g., emotion recognition).<sup>10</sup>

Similarly, this entanglement implies considerable contradictions with data protection principles, such as data minimization and purpose limitation. In other words, snippets may reveal more data than is necessary for a given purpose (e.g., text and typing patterns in language processing).

The coexistence of these different attributes is important when determining the sensitivity of speech and text snippets and determining the legal basis to be used. In particular, it would require taking into account overlapping legal categories of data (e.g., data concerning health, biometric data). In turn, this overlap may mandate the performance of risk assessments that consider the complex nature of speech and text snippets, and the different attributes revealed (e.g., biometric and health attributes).

This challenge has become even more relevant after the Court of Justice of the European Union's ruling *OT v Vyriausioji tarnybinės etikos komisija*<sup>11</sup>. In previous years, the question to what extent data protection laws, and in particular the GDPR, offer protection against

<sup>10</sup> See for example the following study commissioned by the European Parliament: "Biometric Recognition and Behavioural Detection" (2021) p.96.

<sup>11</sup> Court of Justice of the European Union, Judgment of 1 August 2022, (*OT v Vyriausioji tarnybinės etikos komisija*), C-184/20, ECLI:EU:C:2022:601: "[...] Article 9(1) of Regulation 2016/679 must be interpreted as meaning that the publication, on the website of the public authority responsible for collecting and checking the content of declarations of private interests, of personal data that are liable to disclose indirectly the sexual orientation of a natural person constitutes processing of special categories of personal data, for the purpose of those provisions

sensitive inferences (Article 9<sup>12</sup>) or remedies to challenge inferences or important decisions based on them (Article 22(3)) has been discussed in legal scholarship. Wachter et al., for instance, have pointed to significant shortcomings in this regard and concluded that individuals are granted little control and oversight over how their personal data is used to draw inferences about them [4]. In the ruling *OT v Vyriausioji tarnybinės etikos komisija*, the Court had the opportunity to illuminate the question whether Article 9 of the GDPR applies in the situation where special categories of personal data are not explicitly made public (more notably, in online declarations of interests by persons working in the public service as required under Lithuanian anti-corruption law), but Internet users may nevertheless infer certain sensitive information about the declarants, including their political opinions or sexual orientation. In other words, the personal data that needs to be published according to the Lithuanian anti-corruption law are not, inherently, sensitive data in the sense of the GDPR. However, it was possible to deduce from the name-specific data relating to the spouse, cohabitee or partner of the declarant certain information concerning the sex life or sexual orientation of the declarant and his or her spouse, cohabitee or partner. The question to be answered by the Court was, consequently, whether data that are capable of revealing the sexual orientation of a natural person by means of thinking (e.g., involving comparison or deduction) fall within the special categories of personal data, for the purpose of Article 9(1) of the GDPR. The Court confirmed the Advocate General’s opinion from December 2021, namely that Article 9(1) must effectively be interpreted as meaning that the processing of special categories of personal data includes publishing the content of the declaration of interests on the website of the controller in question. In other words, the Court interprets the scope of Article 9 of the GDPR to include sensitive inferences, something advocated for by Wachter et al. [4].

Risk assessments may also need to be performed taking into account the impact of the processing on fundamental rights [5]. For example, under European data protection laws, controllers are obliged to carry out Data Protection Impact Assessments. Article 35 of the GDPR mandates such assessment where a type of processing is “likely to result in a high risk to the rights and freedoms of natural persons”. Similarly, Article 7 of the upcoming AI Act expects the European Commission to consider the risks to individuals’ fundamental rights when amending the list of high-risk AI systems. In relation to speech and language technologies, what would these obligations mean for data controllers when considering the principle of non-discrimination and the right to freedom of speech? Would new fundamental rights be necessary (e.g., right to freedom of emotions)?

### 4.3.3 Vulnerable groups

From a legal perspective, special attention must be given to the concept of vulnerability. Under the upcoming AI Act, vulnerability will be introduced under two key provisions. Firstly, the impact on vulnerable individuals or groups is a determining factor to qualify certain AI practices as unacceptable practices. For example, Article 5 prohibits the use of AI systems that exploit “any of the vulnerabilities” of a specific group of persons due

---

<sup>12</sup> Article 9(1) of the GDPR (previously Article 8(1) Directive 95/46) provides for the prohibition, inter alia, of processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of data concerning a natural person’s sex life or sexual orientation. According to the heading of those articles, these are special categories of personal data, and such data are also categorized as “sensitive data” in recital 34 of Directive 95/46 and Recital 10 of the GDPR.



to their age, physical or mental disability when such use would distort their behavior in a manner that causes or is likely to cause physical or psychological harm. Similarly, the use by public authorities of AI systems to evaluate or classify the trustworthiness of individuals based on social behavior, known or predicted personal or personality characteristics are also prohibited, under certain conditions, when it leads to detrimental or unfavorable treatment of certain individuals or groups.

Additionally, the concept of vulnerability is also used as a factor to be assessed by the European Commission when amending the list of high-risk AI systems. Under Article 7, the European Commission needs to consider:

- the extent of harm or adverse impact of AI systems in terms of intensity and ability to affect a plurality of persons and
- whether impacted persons would be in a vulnerable position, particularly due to an imbalance of power, knowledge, economic or social circumstances, or age.

At the same time, the concepts of vulnerability and vulnerable groups in relation to language technologies raise many questions from an inter-disciplinary perspective.

When speech recognition or natural language processing systems are utilized on a broad scale, these systems at some point will interact with individuals from the so-called vulnerable groups. This broad term typically includes humans with conditions that require special consideration, both in a technical and legal dimension. We can distinguish three types of vulnerable groups of relevance here:

1. individuals with special characteristics of their voice or language,
2. individuals that are not themselves able to utilize their human rights, and
3. individuals that belong to discriminated groups due to special personal characteristics like sexual orientation, ethnicity, or religious or political position.

In the first group, people with speaking issues like stuttering, aphonia, or amnesic aphasia clearly become relevant. The so-called “Doddington zoo” effect [6] also means that some people’s voices are more easily identifiable than others for reasons that cannot be traced back to a specific characteristic. As discussed previously, AI-based speech recognition works with training based on a large set of speech examples, which may or may not have contained people with these specific conditions. If present, the trained AI might be able to cope with (and hide) the specific type of speech characteristics, but if the training dataset did not contain such examples, it might work less well when confronted with speech or language examples from such individuals. Hence, one challenge lies in the proper and non-biased selection of training data, as inclusion of all possible speech- or language-specific abnormalities in the training dataset tends to raise discriminatory real-world issues in itself. As an example, consider an advertisement explicitly asking for stutterers to join a training dataset recording. The resulting dataset would be biased towards favoring stutterers to other speech issues, and the real-world discriminatory effects of such an advertisement could be socially challenging as well.

The second group requires close attention, especially from the legal point of view. Transfer of self-responsibility to another human is a severe and highly sensitive issue, and should only be done in cases that have no alternative. Children are especially vulnerable in this case, as they cannot oversee the consequences of their actions sufficiently, so their parents or legal guardians have to approve decisions or even make decisions themselves for the children. In terms of speech-based interaction technology, this dependency of a child towards its custodian makes the former especially vulnerable, as audio surveillance of sleeping babies is a common and mostly socially accepted scenario. However, this raises a lot of open issues when it comes to questions of secondary use of the voice data created by children, e.g., towards advertising or psychological analysis by third parties – especially in the long term, when these children grow up to be adults of the same personality.



Another example of the second group type is people with diseases like dementia or mental disorders. Even if these may at some point decide to e.g., utilize smart speakers in their homes, or consent to having their language in a social media chat app get analyzed by a research institution, this decision may not stay aware to them. Hence, subsequently, when confronted with the ongoing voice surveillance of the smart speaker, or receiving the feedback from the research institutions, such individuals may suffer from severe trauma. On the other hand, availability of such technical surveillance or assistance systems might be very beneficial towards these individuals, especially for those also suffering from physical deficiencies like inability to type or utilize other input devices for a computer.

The third group is special in a large variety of possible ways, ranging from sexual orientations that are considered illegal in some countries of the world to social discrimination or even physical frays based on skin color, nationality, or political opinions expressed. In all of those cases, speech and language processing systems to some extent may be able to identify such conditions, based on what was said or how it was said in specific contexts (e.g., lie detection when confronted directly).

In general, belonging to a vulnerable group is no explicit act, and the definitions of what substantiates a vulnerable group differ largely.

What is common to them is that speech and language processing systems have to be designed in a way that they are either reliably agnostic to these conditions or consider them appropriately in the design and behavior of the system in consideration. Here, privacy-enhancing technologies may help, and should be considered wherever possible.

#### 4.3.4 Confidentiality vs. duty to rescue

In some situations, the users' right to privacy may conflict with the voice technology company's legal requirements. For example, if the voice technology company collects speech or text data suggesting that a crime (e.g., child abuse) or a life-threatening danger (e.g., heart attack) has taken place, should it report it to the relevant authority, thereby violating the user's privacy? Is it enough to report cases that have been incidentally found or should the company be required to automatically analyze the data to find all possible cases and have them screened by a human operator, which is a form of systematic surveillance? When answering these questions, it is important to realize that legal requirements regarding "duty to rescue" vary from one jurisdiction to another.<sup>13</sup> In most jurisdictions under civil law (Europe, Latin America) and in some US states, it is a legal duty for citizens to assist in such cases unless this would put them in danger, with some exceptions (e.g., if the citizen is a priest or a lawyer hearing a person confess a crime, the confidentiality obligation is stronger). The duty to rescue does not apply to companies in these jurisdictions nor to citizens or companies in other jurisdictions, which implies that such cases can be reported but it's not an obligation. Nevertheless, some companies have been requested by law enforcement agencies to automatically screen for, e.g., child pornography in personal image data. This raises three open questions. From a societal point of view, should companies be requested, allowed, or forbidden to perform large-scale automatic screening in the speech and text data they collect? If this is requested or allowed, what should be the territorial extent (e.g., would it apply to a European company processing data from an American citizen) and which legal safeguards should be put in place to preserve fundamental human rights regarding censorship (what can or cannot be uploaded) and massive surveillance? Also, from a technical point of view, could this screening be performed on-device in a privacy-preserving way?

---

<sup>13</sup>[https://en.wikipedia.org/wiki/Duty\\_to\\_rescue](https://en.wikipedia.org/wiki/Duty_to_rescue)

## References

- 1 Catherine Jasserand. Legal nature of biometric data: From generic personal data to sensitive data. *European Data Protection Law Review*, 2(3):304, 2016.
- 2 Els Kindt. Having yes, using no? About the new legal regime for biometric data. *Computer Law & Security Review*, 34(3):523–538, 2018.
- 3 Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur. Probing the information encoded in x-vectors. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 726–733. IEEE, 2019.
- 4 Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review*, 2019(2), 2019.
- 5 Dara Hallinan and Nicholas Martin. Fundamental rights, the normative keystone of DPIA. *European Data Protection Law Review*, 6(3):178–193, 2020.
- 6 George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Technical report, DTIC Document, 1998.

## 4.4 Privacy attacks

Abdullah Elbi (KU Leuven, BE) [abdullah.elbi@kuleuven.be](mailto:abdullah.elbi@kuleuven.be)

Anna Leschanowsky (Fraunhofer IIS – Erlangen, DE) [anna.leschanowsky@iis-extern.fraunhofer.de](mailto:anna.leschanowsky@iis-extern.fraunhofer.de)

Pierre Lison (Norsk Regnesentral – Oslo, NO) [plison@nr.no](mailto:plison@nr.no)

Andreas Nautsch (Avignon Université, FR) [andreas.nautsch@univ-avignon.fr](mailto:andreas.nautsch@univ-avignon.fr)

Olga Ohrimenko (University of Melbourne, AU) [oohrimenko@unimelb.edu.au](mailto:oohrimenko@unimelb.edu.au)

Laurens Sion (KU Leuven, BE) [laurens.sion@kuleuven.be](mailto:laurens.sion@kuleuven.be)

Marc Tommasi (University of Lille, FR) [marc.tommasi@univ-lille.fr](mailto:marc.tommasi@univ-lille.fr)

License © Creative Commons BY 4.0 International license

© Abdullah Elbi, Anna Leschanowsky, Pierre Lison, Andreas Nautsch, Laurens Sion, Marc Tommasi

### 4.4.1 Context and motivation

One way to assess the strength of privacy-enhancing techniques (and the data protection they provide) is to conduct so-called *privacy attacks*. In our context, a privacy attack is a process which, given a particular input or model, seeks to uncover personal data that should be or should have been concealed. Privacy attacks can be employed as part of privacy risk assessments (including Data Protection Impact Assessments) or as an evaluation method in the development of privacy-enhancing techniques.

It is, however, important to stress that privacy attacks can usually only provide lower bounds when it comes to assessing the privacy risk associated with a given output or model. Privacy attacks are by construction not exhaustive and can only explore a limited region of the risk space. In other words, they can only demonstrate the presence of a privacy risk and not their absence. Although we can make assumptions about possible attackers and the background knowledge those attackers may have access to, those assumptions may very well turn out to be invalid. Attackers may also rely on other attack strategies than the ones that have been explicitly tested.

Although the present section focuses specifically on privacy attacks (i.e., attacks designed to uncover personal data), it is worth noting that security attacks (i.e., attacks targeting the confidentiality, integrity, or availability of an IT system) may also lead to privacy breaches. In particular, it has been shown that one can infer the hidden values of a black-box machine

learning model based on requests made on this model, but we consider such attacks as being primarily a security issue, although they might also lead to privacy risks. Another example is poisoning a model, which could lead to adverse decisions or inferences being made regarding another individual.

#### 4.4.2 Core concepts

Let us assume a function  $f$  used to transform some raw data  $D$  into another data or model  $D'$ , as illustrated in Fig. 3. This transformation may correspond to a sanitization/anonymization process or to the training of a machine learning model. Examples of such data include:

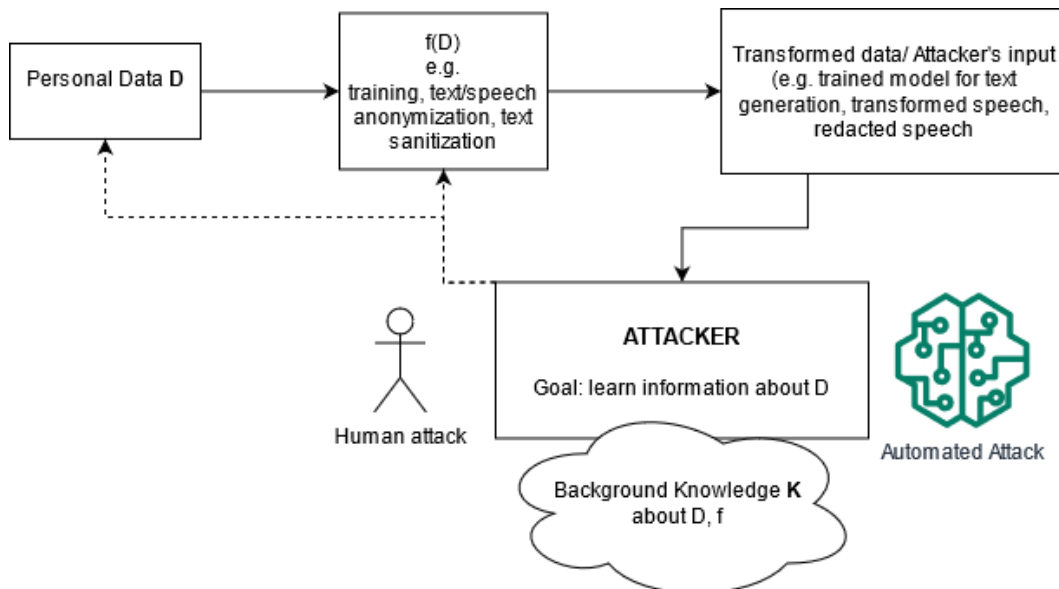
- raw speech recordings (prior to speech anonymization),
- initial text documents (prior to text sanitization),
- documents or speech recordings employed to train a machine learning/natural language processing model.

The result  $D'$  of the transformation  $f(D)$  may take a variety of forms. It may correspond to a “sanitized” or anonymized version of  $D$ , but may also correspond to a trained model. We assume that both the transformation  $f$  and the outcome  $D'$  are known to the attacker (but not the personal data  $D$ ).

We also assume that the raw data  $D$  contains (potentially sensitive) personal data. This ranges from data explicitly stated in the raw data (e.g., phrases or spoken words, location) to implicit data that can be learnt from raw data (e.g., speaker identity, author, emotional state, age, country of origin, gender, background noise including third parties) and metadata. This information may also be sensitive and belong to “special categories” of personal data in the GDPR such as health data, religious beliefs, or ethnic origin.

The goal of the attacker is then to infer some personal data, based on the observation of the transformed data/model  $D'$ , function  $f$ , and background knowledge  $K$ .

The attacker can target different stages of the data processing pipeline, including:



■ **Figure 3** Core concepts of privacy attacks.

- the *training phase*, where the model is trained based on a training dataset which may include personal data,
- the *deployment/operational phase*, where the model is deployed to achieve identified objectives (for example, verifying the identity of individuals in the case of biometric verification).

#### 4.4.2.1 Attack goal

The attacker may seek to retrieve various types of personal data, such as:

- the identity of the speaker (or author of a document),
- the identity of a person mentioned in an audio recording or text document (who may be distinct from the speakers/authors),
- (potentially sensitive) personal attributes associated with those individuals, such as their gender, age, emotional state, or country of origin,
- Whether a person is included/described in the raw dataset  $D$ .

Various mechanisms have been developed to conduct and demonstrate such types of attacks.

#### 4.4.2.2 Attacker role

The attacker role considers whether the attacker has a privileged role in or access to the system. It distinguishes between attackers that are outsiders without any type of access and attackers that are involved in the data processing as a (trusted) third party, or attackers that have insider access to the system. Insider access can be the result of a consumer-service provider relationship (e.g., a voice assistant company storing and analyzing user data), an interpersonal relationship (e.g., a spouse that knows a user's passwords), an authoritative relationship (e.g., an employer or school), and possibly other relationships as mechanisms.

The nature of the attacker role may therefore have an impact on the attacker capabilities, the attacker's knowledge on the system, or the attacker's background knowledge we describe below.

#### 4.4.2.3 Attacker capabilities

The attacker's capabilities can be different while having the same intention. We distinguish between those that can manipulate the raw data, or raw data transformation to those that cannot. For example, the former can intervene by contributing raw data into the processes (e.g., inject data) or manipulate the data transformation process (e.g., affecting the training code to memorize the data).

Attacker capabilities should also take into account the amount of computational resources the attacker has access to. Greater computational resources can increase the threat an attacker poses.

#### 4.4.2.4 System knowledge of the attacker

Different attackers may have different degrees of knowledge about the system of interest. Some attackers may not have any information about the system at all and only have access to the outputs of the system. Others may have full access to the system and its source code and can leverage this information as part of their attacks. In the case of an attack in the training phase of a model using federated learning, the attacker may have access to a set of gradients, a set of models or a set of training losses.

#### 4.4.2.5 Background knowledge of the attacker

Attacker models typically rely on assumptions regarding the information or tools that may be available to an adversary seeking to uncover the information that should be protected in the research or deployment phases of the machine learning tools. For instance, an adversary seeking to determine the identity of the speaker of a given speech segment may be assumed to have access to audio recordings of various potential candidates. Similarly, an adversary seeking to find out the identity of a person mentioned in a sanitized text will typically have access to public information sources (like web data) about potential individuals. This background knowledge may take various forms, and may also correspond to machine learning models or computer software. Attack models should take into account as much background knowledge as possible to ensure the attacks are sufficiently strong.

#### 4.4.2.6 Attack mechanisms

Privacy attacks can be implemented through a range of possible techniques. Attacks can be conducted through automated inferences, and/or by humans seeking to uncover some unintended information based on various knowledge sources. For example, the attacker can train a model on an auxiliary dataset, resembling the structure of the model it is attacking. The adversary may also have access to multiple versions of the model (e.g., original and updated/fine-tuned model) and try to extract information by accessing the two in parallel. For such automated attacks, one also needs to make assumptions regarding the computational power that we expect to be available to a motivated attacker (see attacker capabilities).

#### 4.4.2.7 Who conducts the attack and how it is evaluated

A final dimension to consider is who is practically responsible for designing and conducting the attack. This can be the organization developing the system, a third party, or a data protection authority. Privacy attacks also have a success rate that varies depending on attributes or personal or protected information of an individual (e.g., members of a vulnerable group maybe more susceptible than others). To this end, privacy attacks should be carried out on wide data distribution.

### 4.4.3 Concrete example attacks

We list the following example attacks and information they can extract. Some of these attacks should be adapted to speech and text (e.g., what does membership mean in this context) as they were first proposed on tabular data.

- *Membership inference attack*: The adversary tries to find out whether a certain data record (e.g., a data record can be a piece of text contributed by a user) was present in the personal data fed to the transformation function  $f$  [1]. In relation to speech or text, this may correspond to whether a certain person contributed to the corpus with text or voice that the attacker holds and where participation in the dataset could be considered personal or sensitive (e.g., written descriptions of patients' medical condition). A more general attack could be a *presence attack* where the attacker tries to infer if a given person has contributed to the dataset from examples of his voice or his writings whether or not these examples are present in the dataset. These attacks for text models can be used to see if user's data was used in training.
- *Re-identification attack*: The adversary attempts to determine an individual's identity. For example, in the case of transformed or anonymized text whether an attacker can

determine the identity of the person or narrow down to a small enough sample of users. This is an instance of attribute inference where the attribute is an identifier of a person.

- *Attribute inference*: Given a trained model as the output of the transformation function  $f$ , and some information about non-sensitive attributes of the individual whose privacy we want to attack, the adversary tries to reconstruct the sensitive attributes. Applicability of such an attack to speech and language data is questionable due to 1) the non-tabular nature of the data, and 2) the potential lack of agreement on what sensitive attributes of text and speech are, also taking into an account their availability in the training data (e.g., labeling such attributes beyond those “easily” determinable, such as age or occupation).
- *Data extraction (aka model inversion attack)*: The adversary tries to extract verbatim training data from a trained model [2, 3].
- *Update data extraction*: The adversary tries to extract the data that was used to update or fine-tune a model based on interaction with several models [4].

Liu et al. [5] provide a survey of privacy attacks on machine learning models.

#### 4.4.4 Open questions and challenges

As mentioned earlier, privacy attacks can only explore a limited number of possible attacks. Due to this empirical nature of the evaluation, we believe that is important to vary the types of attackers, their knowledge, and to develop metrics and guarantees that can be given based on the attack success rate, for example, in terms of confidence intervals.

It remains unclear how to conduct a privacy attack on arbitrary types of language data, in particular for the case when the person to protect is not the author of the speech recording or text document, but corresponds to a third-party mentioned in those. An interesting strategy would be to use reinforcement learning or a GAN-inspired framework to help the attack model learn if personal data is leaked through the transformation  $f$ . As re-identification often proceeds in a sequence of reasoning steps, the use of “chain of thought” prompting [6] also constitutes a promising approach to re-identify individuals from speech or text data.

Most automated attack mechanisms have been developed by researchers. We believe creating a framework for automatically generating attacks would help streamline the process and potentially identify new attacks. This framework could for instance take the form of an open-source library of attacks/implementation, structured according to the attack mechanism, analogously to MITRE’s CAPEC<sup>14</sup>. Alternatively, one could also make available a standard API to test your model against privacy attacks.

#### References

- 1 Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P)*, pages 3-8, 2017.
- 2 Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 587–601, 2017.
- 3 Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.
- 4 Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. Analyzing information leakage of

<sup>14</sup>Common Attack Pattern Enumeration and Classification, MITRE, <https://capec.mitre.org>

updates to natural language models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 363–375, 2020.

- 5 Chao Liu, Xin Xia, David Lo, Cuiyun Gao, Xiaohu Yang, and John Grundy. Opportunities and challenges in code search tools. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021.
- 6 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

## 4.5 Privacy enhancing technologies

*Martine De Cock (University of Washington – Tacoma, US) mdecock@uw.edu*


*Zekeriya Erkin (TU Delft, NL) z.erkin@tudelft.nl*

*Simone Fischer-Hübner (Karlstad University, SE) simone.fischer-huebner@kau.se*

*Meiko Jensen (Karlstad University, SE) meiko.jensen@kau.se*

*Dietrich Klakow (Saarland University – Saarbrücken, DE) dietrich.klakow@lsu.uni-saarland.de*

*Francisco Teixeira (Instituto Superior Técnico – Lisbon, PT) francisco.s.teixeira@tecnico.ulisboa.pt*

**License**  Creative Commons BY 4.0 International license

© Martine De Cock, Zekeriya Erkin, Simone Fischer-Hübner, Meiko Jensen, Dietrich Klakow, Francisco Teixeira

Privacy-enhancing technologies (PETs) provide technical building blocks for achieving privacy by design and can be defined as technologies that embody fundamental data protection goals [13] including the goals of unlinkability, intervenability, transparency and the classical CIA (confidentiality, integrity, availability) security goals by minimizing personal data collection and use, maximizing data security, and empowering individuals.

The privacy by design principle of a positive sum for speech and language technologies should enable users to benefit from the rich functions of these technologies while protecting the users’ privacy at the same time. The fundamental question is how to achieve privacy by design for speech and language technology without hampering the services. To achieve this goal, different PETs exist that can be utilized for this purpose. Below, we first discuss what type of personal data are accessible via speech and text and should be the target of protection by PETs. Then, we provide an overview of PETs that can provide protection and discuss their limitations and challenges that arise when used for speech and language technologies.

### 4.5.1 Possible private content

When discussing PETs’ capabilities and limitations, the first important dimension is the possible private content contained within speech and text. The different steps in the processing pipeline and the different levels of information contained might make the use of different PETs necessary.

For this we can look at the standard processing pipeline of a speech signal. We start by recording speech with one or more microphones; this is usually followed by some form of audio pre-processing, such as denoising, enhancement or separation. The life cycle of the speech signal then becomes dependent on the target application. If the target application is related with purely acoustic characteristics of the signal, speech can be fed directly into the



corresponding processing pipeline. If the end goal is to analyse and process text, then speech must first be transcribed either by automatic or manual means. Of course, text can also be produced by other means and will also be considered here.

Below we provide a summarized categorization of the levels of information contained within speech and text, to allow for a more detailed analysis on the necessity and application of PETs in the context of the processing of speech and text.

#### 4.5.1.1 Speech

Being both a means for communication and a biometric signal, speech is formed by multiple dimensions, each containing different levels and types of information.

From a technological point of view, to analyze speech, we need to first record it. This means that a recording will contain voice, environmental acoustics, and information about the recording setup.

By voice, we mean any acoustic phenomenon produced by an individual's vocal tract, which can include verbal and non-verbal dimensions. The verbal dimension corresponds to lexical or phonetic content, that has a direct correspondence to language and communication. The non-verbal dimension can be further divided in two sub-dimensions, one that is non-verbal but communicative, and a second that is non-verbal and non-communicative. More formally, we within voice we have:

- Lexical information – also called verbal [3] or linguistic content [30], represents phonetic content associated with language that is meant to be communicative.
- Paralinguistic information [22] – acoustic factors related with communication which provide, consciously or otherwise, additional information to the listeners (e.g., emotion and intent, accent).
- Extra-linguistic Information[22] – acoustic factors not directly related to communication, corresponding to speaker characteristics [3] which arise from mental, biological, and physical traits of the speaker, such as the speaker's age, gender and health status. This information may also be characterized as paralinguistic [30].

Environmental acoustics include other background sounds, which may provide information about the speaker's location, surroundings, and communication context. Moreover, the type(s) of microphone(s) or device(s) used to create the recording will also be reflected in the recording itself, making it possible to even identify the specific device used to create it.

#### 4.5.1.2 Text

Text can contain private information irrespective of whether it is produced from speech or typed directly. Possible sources for text from speech can be transcription tasks, proper dialog systems or any human-to-human communication that is transcribed after the prime use case (e.g., speech messages sent over WhatsApp). Possible sources of text are medical reports, legal course rulings, messages or e-mails.

Information derived from writing style is given away involuntarily. Here are some examples:

- Gender is often given away by the use of short function words e.g., the usage of “would” and “may” [12].
- Age: like with gender, word usage can give away the age of the writer. This could also be captured in a very single language model [12].
- Mother tongue: some studies have shown that the mother tongue is given away e.g., by the usage (or non-usage) of determiners.



- Literacy: statistical features measured by stylometric measurements are also detectable and can be grouped by
  - Length-based measurements: average length of words, sentence length in words and characters;
  - N-grams: the frequency of groups of two, three or more words;
  - Lexical richness measurements: values scored by frequency of the vocabulary and all occurred tokens (words and punctuations), in detailed, e.g., Giraud's  $R$ , Herdan's  $C$ , Dugast's  $k$ , Maas'  $a^2$ , Tuldava's  $LN$ , Brunet's  $W$ , Summer's  $S$ , Horoné's  $H$ , Sichel's  $S$ , Michéa's  $M$ , Entropy, Yule's  $K$ , Simpson's  $D$ , and Herdan's  $V_m$ ;
  - Readability and formality scores: values scored by syllables, frequency of often-used words, counts of numeral words given by, e.g., Flesch's reading ease formula [9], Flesch-Kincaid formula [10], Dale-Chall scored by a list of 3000 words [7], McLaughlin's SMOG formula [24], FORCAST formula (US military, 1973), Gunning fog index [11], Automated readability index (ARI) [31], or Heylighen formality score [15];
  - Syntactical features: measured by scores depending on dependency and/or constituency grammar parsers, e.g., occurrences of elements from parse trees;
  - Semantic features: scores measured in connection with word wise request on semantic networks, e.g., WordNet, GermaNet (only for German language) or biomedical terminology networks, e.g., UMLS and/or SNOMED CT.
- Information derived from semantics: the prime purpose of text is to convey an explicit message. The open problem is to decide what the core content of a text is and which part should be protected. For practical purposes so-called named entities are often used as proxies. Those are person names, organization names or locations. Often dates or monetary amounts are also detected by the same software.

#### 4.5.2 Training or inference

In machine learning applications for speech and language technology, personal data is typically used in two different stages.

*Training.* During the model development phase, a model is induced from training instances. The training instances consist of personal speech and/or text data from users that may be labeled (supervised learning) or not (unsupervised learning). One may need to protect:

- the training instances themselves [14] (this is sometimes referred to as input privacy),
- the resulting trained model, as the model itself can leak information about the training instances [4] (output privacy).

Furthermore, one may also want to protect model integrity, for instance to prevent malicious actors from poisoning the training data to deliberately hinder the model's performance or to make it more vulnerable to attacks [18].

*Inference.* After a model is trained, it can be deployed and used in the inference phase during which a trained model is used to deliver a service, such as emotion recognition from speech. One may need to protect [28, 34]:

- the model that is used to make the inferences – such as the emotion recognition model – as the model itself can leak information about the training instances (input privacy),
- the query instance – such as the snippet of speech to be classified – as it reveals information about the speaker (input privacy),
- the result of the inference – such as the inferred emotion – as it leaks information about the model and the query instance (output privacy).

Beyond preventing leakage of personal user data from the trained model, one may also want to protect the intellectual property (IP) of the owner/developer of the model. The latter

is particularly relevant when the trained model constitutes a competitive advantage, or in security applications such as spam or hate speech detection, where knowledge of the model would help adversaries to develop strategies for evading detection.

Orthogonal to the above, machine learning is also used for privacy-preserving release of speech and text data, in particular to recognize and mask personal data in text documents (such as court cases and medical records) and in speech signals.<sup>15</sup>

### 4.5.3 Training/inference locally, centrally or collaboratively

A relevant issue for both the training and inference phases is how the data is distributed, and where and by whom it is processed.

*Training* of machine learning models is traditionally done in a *central* way, assuming that one entity has access to all the training instances and uses it to induce a model. In many applications however, the data naturally originates from multiple entities who may be capable of each training their own model *locally*, or who may want to *collaborate* in training a joint model however without sending their raw data to each other or to a central entity.

*Inference*, namely classifying or scoring a new query instance with a trained machine learning model, inherently involves two entities, namely the model owner and the query instance owner (a.k.a. the user). Inference can be done *locally* by the model owner (which requires the user's query to be sent to the model owner); or locally by the user (which requires the model to be deployed on the user's end); or jointly by both in a *collaborative* manner. It can also be *outsourced* to yet another entity in the cloud.

### 4.5.4 The privacy enhancing techniques

Hoepman [16] introduced eight strategies for privacy: minimize, hide, separate, aggregate, inform, control, enforce and demonstrate. In the following, we use these strategies as a means for classifying existing PETs for speech and language technologies.

- “Minimize”: This strategy enforces the basic privacy principle of data minimization by limiting the amount of personal data that is processed, e.g. by avoiding data collection from the start of by using anonymization or pseudonymization techniques.
  - Pseudonymization can be used to render text pseudonymous by replacing identifiers with pseudonyms. Text pseudonymization is often used in the context of clinical text de-identification, where for instance patient name are to be replaced by pseudonyms (see e.g. [26] for clinical text pseudonymization based on deep learning). Another example for an automated tool that recognizes and pseudonymizes privacy-relevant text parts in private communication (email) is provided by [8].
  - Typical anonymization techniques use data generalization or suppression (e.g. k-anonymization or variants) or data perturbation by adding statistical noise to aggregated data (e.g., differential privacy).
  - For text processing, k-anonymity has a number of applications. Summarizing text while hiding identifying attributes is a simple approach used in practice. In natural language processing, there are extensions for k-anonymity, e.g., t-plausibility[2] for generalizing words to desensitize text or c-sanitized [29], an information theoretic approach based on finding sensitive terms that have a high mutual information with an entity. These terms can be taken alone or in combination. After, they can be redacted or replaced with non-sensitive substitutes.

<sup>15</sup><https://www.voiceprivacychallenge.org/>

- Differential privacy (DP) is a data perturbation technique that can be applied locally on the user’s device (local DP) or centrally for federated learning. Local DP approaches exist that are rewriting a text until a certain privacy guarantee (epsilon) is reached (see, e.g., [17]).
- Further minimization methods that are specific to speech include voice anonymization techniques (see, e.g., [6, 35]).
- A systematic review of deep learning methods for privacy-preserving natural language processing, including data minimization PETs categorized into de-identification (corresponding to pseudonymization) methods, anonymization methods and differential privacy methods, is provided by [32].
- “Hide”: This strategy aims at disallowing access to personal data in plain view or restricting access to only authorized users, meaning that precautions such as transmitting data securely and storing data in the encrypted form should be considered. Furthermore, it would also be difficult for adversaries to observe the data flow by deploying mix-nets, onion routing and similar technologies. Cryptographic tools and techniques such as homomorphic encryption (HE – protecting input data and the result of computation) or functional encryption (FE – which in contrast to HE only protects the input data while the computation result is made available in cleartext), as well as secure multiparty computation (MPC) and secure distance-preserving hashing (DPH) or randomized binarization also provide data hiding.
- “Separate”: This strategy includes data or process separation by processing data including metadata in a distributed manner if possible. The data should be kept in a separate form: tables and datasets should be divided into different tables and if possible different datasets and locations.
  - A good example of data separation is using additive secret sharing, a.k.a. multi-party computation, where data is split into random parts and each part is kept in different servers. In this way, the attacker cannot deduce any meaningful information about the data unless she has a certain number of the shares. Secret data shares can still be processed using MPC techniques.
  - Federated learning (FL), which is used for many natural language processing applications, is also based on separation, but is not sufficiently privacy-preserving, as still information can leak. For this reason, FL is more and more used in combination with other PETs, such as DP and HE or MPC.
  - Slicing is another example for enhancing speech privacy via separation. After speech is anonymized, the speech signal is split into small chunks when storing the data such that it becomes harder for an adversary to re-identify the original speaker [23].
- “Aggregate”: This strategy is applying data aggregation for concealing data. It is the step where the amount of personal information is restricted at a group level. By doing so, it is difficult for the adversary to identify a certain individual within the group. Data aggregation via statistics or building machine learning models is however not sufficiently protecting data, as personal information could leak via inference attacks, i.e., via correlation of statistics, and therefore complementary privacy-protecting measures such as DP are needed. To some extent, the use of synthetic data instead of data taken from real human individuals also falls into this category, as the typical characteristics of the real-world data are derived, in aggregated form, and utilized to generate randomized new data that follows the same statistical distribution. Hence, on an aggregated level, the original information on characteristics of the real-world data is still contained in the synthetic data, just on an aggregated level.

- “Inform”: This strategy aims at providing transparency to data subjects when their data are processed. Different types of Transparency enhancing tools (TETs) exist.
  - Ex ante TETs can provide transparency before personal data is disclosed/processed for enabling informed decisions/consent. Privacy product labels on packages, e.g., for IoT (Internet of Things) or natural language processing products, can for instance inform customers about privacy practices when products are purchased (see also [19, 27]). Concepts for usable policies, e.g. as part of consent forms, multi-layered policies, automated policy assistants exist in general but have not specifically been applied for speech and natural language processing. Ex ante TETs explaining how AI and automated decision making is in principle working is still an area of research.
  - Ex post TETs are providing transparency about how data have been processed. Also ex post TETs for explainable AI, which are explaining how decisions have been made, e.g., via attention maps or feature selection, are a current research area. Other general ex post TETs include privacy dashboards for displaying what data has been processed by whom, for tracking data usages and flows or for data export, as well as privacy notification tools (e.g., informing about breaches or risks) (see [25] for an overview).
- “Control”: This strategy should provide data subjects with control over their data. Different intervenability tools have been researched or developed for enhancing control, even though they have been hardly applied to speech and natural language processing yet. These tools include for instance:
  - privacy policy tools/languages exist for negotiating privacy policies between data subjects and data controllers,
  - privacy dashboards which allow data subjects to conduct or request (from the data controllers) data deletions or corrections or data export for data portability,
  - tools for data subjects to object to automated decision making,
  - consent management tools which enable data subjects to provide and to easily revoke any time informed consent.
- “Enforce”: This strategy should guarantee that a privacy policy that complies with data protection/privacy laws is enforced. Access control systems, e.g., based on attribute or role-based access control, can provide means for technically enforcing privacy policies.
- “Demonstrate”: This strategy requires the controller to demonstrate compliance and enable accountability. General approaches and tools for supporting this strategy include:
  - logging and auditing tools and privacy intrusion detection systems which allow to detect privacy breaches as a means for making attackers accountable and to demonstrate compliance,
  - privacy certification of PETs in regard to their privacy functionality and assurance by independent certification bodies as a means for demonstrating compliance – the certification results could be (certified) privacy seals (e.g., EuroPrise seal<sup>16</sup>) that mediate privacy levels of products to users/customers,
  - consent management tools which allow data controllers to easily keep proofs for the data subjects that consent has been provided.

#### 4.5.5 Challenges for PETs

In general, PETs are very valuable tools for implementing privacy into speech and language technology. However, there also are several challenges to consider in the design phase:

<sup>16</sup> <https://www.euprivacyseal.com/EPS-en/Home>

- Difficulty to determine all personal attributes: it is hard to guarantee that all possible different types of personal information from an input are removed or protected by data hiding or access control, because we simply don't know what information might actually be contained in an audio or voice snippet.
- Privacy-performance tradeoffs: PETs, especially homomorphic encryption or MPC, may have severe efficiency tradeoffs, which may not be acceptable if fast responses are needed.
- Privacy-utility tradeoffs: data minimization, aggregation, and data hiding may conceal information or use data perturbation by adding statistical noise, which creates utility tradeoffs.
- Tradeoffs between privacy protection goals: for machine learning models, there is also a tradeoff between protection against member inference attacks and fairness of decision making [5]. Moreover, data minimization and transparency are privacy protection goals that are typically in conflict with each other.
- Usability: PETs based on “crypto-magic” operations are often counterintuitive to users and hard to comprehend and trust.
- The privacy guarantees of PETs rely on an attacker model and its assumptions: for instance, secret sharing and MPC require multiple non-colluding entities that act as independent organizations.

In general, it is a challenge to select and configure the right combination of PETs for addressing privacy trade-offs mentioned above, for sufficiently protecting all personal data and metadata items, and for fulfilling legal requirements posed by the GDPR and its privacy by default and by design principle, the AI act or other regulations.

Beyond these general issues, each specific technique or strategy implementation approach has its very unique challenges when applied to speech and language data, some of which are listed next:

- *Minimize:*
  - K-anonymity is not straightforward to apply on speech data. For example, t-plausibility [2] has strong assumptions that may not be realistic, as they lead to a too high utility loss.
  - For redacting text, deployment heavily relies on the type of the data: while it is possible to redact text for court cases, it might cause problems to do so for medical data. For c-sanitized, the computation of mutual information is not easy and for instance relies on google searches. For text redaction in eHealth application, a tradeoff between privacy and safety and accountability needs to be taken into consideration (see [1]).
  - For redacting speech, deployment may rely on manual redaction, or on auxiliary technologies that allow the transcription of the speech signal, or that perform keyword spotting.
  - Local DP, which is used for approaches of text re-writing, usually requires a high epsilon for still retaining sufficient data utility, and thus cannot provide very good privacy guarantees. For central DP applied for federated learning, there is no confidentiality of the data against the central aggregator that needs to be trusted. (see also discussion of limitation of DP for natural language processing in [21]).
- *Hide:*
  - The interactive nature of MPC might require bandwidth requirements and involvement of the parties to be online for processing. Particularly, in the case of malicious security model, the overhead introduced will be a significant performance burden overall. While this may be acceptable for training of MDATAL models, it can be problematic in inference applications where the responsiveness of the system (i.e., short response

times) matters. MPC does not inherently require adaptations to the system, meaning that for already trained models, no utility loss is introduced.

- For HE computational burden is usually very high, but is supported by the server. Bandwidth costs are relatively low. HE also has limitations in terms of the type and amount of supported operations, which may make the implementation of certain technologies infeasible. The adaptations required by HE can cause utility loss.
- Functional encryption has limitations in the type of operations that may be performed, and also suffers from performance restrictions. Applications of FE are different from HE as the processing party in FE is the party who learns the result.
- *Separate:*
  - As we discussed above, FL is not fully privacy-preserving. Information from the clients may leak to the central server, and to other clients. For this reason, FL is more and more used in combination with other PETs, such as DP and HE/MPC.
  - Another problem of FL is that it requires the clients to have sufficient training data. This is especially challenging if one expects that the training data is annotated, i.e., for supervised learning. While unsupervised FL with text data has been studied, to the best of our knowledge, unsupervised FL for speech is underexplored.
- *Aggregate:* [33] have recently shown that synthetic data does not provide a better trade-off between privacy and utility than traditional anonymization techniques, which is additionally even harder to predict than for traditional techniques.
- *Inform:*
  - Explainable machine learning: There is an inherent tension between privacy and explainability, as providing the user with an explanation of how a machine learning system reached an outcome inevitably entails leaking some information about the machine learning model and possibly the underlying training data.
  - Explaining the protection functionality of PETs remains a usability issue. For instance, [20] revealed several common misconceptions that lay users develop if confronted with metaphors for differential privacy that are commonly used by media outlets.
  - In general, there is a lack of TETs and of control tools for speech and language technology that users can use in practice for exercising their data subject rights.
- *Control:* Intervenability tools that major natural language processing providers, such as Google and Apple, do not support fine-grained controls in the form of deletions or corrections and only limited insights or controls for derived/inferred data (e.g., data portability is usually not provided for inferred data).
- *Enforce:* Defining fine-grained access control policies that also sufficiently protect not only content data but also metadata that can be inferred from speech and language processing, remains a research challenge.

## References

- 1 Ala Sarah Alaqra, Simone Fischer-Hübner, and Erik Framner. Enhancing privacy controls for patients via a selective authentic electronic health record exchange service: qualitative study of perspectives by medical professionals and patients. *Journal of Medical Internet Research*, 20(12):e10954, 2018.
- 2 Balamurugan Anandan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si. t-plausibility: Generalizing words to desensitize text. *Trans. Data Priv.*, 5(3):505–534, 2012.
- 3 Anton Batliner, Simone Hantke, and Björn Schuller. Ethics and good practice in computational paralinguistics. *IEEE Transactions on Affective Computing*, 13(3):1236–1253, 2020.



- 4 Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.
- 5 Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE, 2021.
- 6 Alice Cohen-Hadria, Mark Cartwright, Brian McFee, and Juan Pablo Bello. Voice anonymization in urban sound recordings. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2019.
- 7 Edgar Dale and Jeanne S Chall. A formula for predicting readability: Instructions. *Educational Research Bulletin*, pages 37–54, 1948.
- 8 Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. Code alltag 2.0—a pseudonymized german-language email corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4466–4477, 2020.
- 9 Rudolf Flesch. A readability formula in practice. *Elementary English*, 25(6):344–351, 1948.
- 10 Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221, 1948.
- 11 Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.
- 12 Yaakov HaCohen-Kerner. Survey on profiling age and gender of text authors. *Expert Systems with Applications*, page 117140, 2022.
- 13 Marit Hansen, Meiko Jensen, and Martin Rost. Protection goals for privacy engineering. In *2015 IEEE Security and Privacy Workshops*, pages 159–166. IEEE, 2015.
- 14 Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.
- 15 Francis Heylighen and Jean-Marc Dewaele. Formality of language: definition, measurement and behavioral determinants. *Interne Bericht, Center “Leo Apostel”, Vrije Universiteit Brussel*, 4, 1999.
- 16 Jaap-Henk Hoepman. Privacy design strategies. In *IFIP International Information Security Conference*, pages 446–459. Springer, 2014.
- 17 Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. Dp-rewrite: Towards reproducibility and transparency in differentially private text rewriting. *arXiv preprint arXiv:2208.10400*, 2022.
- 18 Bargav Jayaraman, Esha Ghosh, Huseyin Inan, Melissa Chase, Sambuddha Roy, and Wei Dai. Active data pattern extraction attacks on generative language models. *arXiv preprint arXiv:2207.10802*, 2022.
- 19 Johanna Johansen, Tore Pedersen, Simone Fischer-Hübner, Christian Johansen, Gerardo Schneider, Arnold Roosendaal, Harald Zwingelberg, Anders Jakob Sivesind, and Josef Noll. A multidisciplinary definition of privacy labels. *Information & Computer Security*, (ahead-of-print), 2022.
- 20 Farzaneh Karegar, Ala Sarah Alaqra, and Simone Fischer-Hübner. Exploring user-suitable metaphors for differentially private data analyses. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 175–193, 2022.
- 21 Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. Differential privacy in natural language processing: The story so far. *arXiv preprint arXiv:2208.08140*, 2022.
- 22 John Laver, editor. *Principles of Phonetics*. Cambridge University Press, 1994.
- 23 Mohamed Maouche, Brij Mohan Lal Srivastava, Nathalie Vauquier, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. Enhancing speech privacy with slicing. In *Interspeech 2022-Human and Humanizing Speech Technology*, 2022.

- 24 G. Harry Mc Laughlin. Smog grading – a new readability formula. *Journal of Reading*, 12(8):639–646, 1969.
- 25 Patrick Murmann and Simone Fischer-Hübner. Tools for achieving usable ex post transparency: a survey. *IEEE Access*, 5:22965–22991, 2017.
- 26 Jihad S. Obeid, Paul M. Heider, Erin R. Weeda, Andrew J. Matuskowitz, Christine M. Carr, Kevin Gagnon, Tami Crawford, and Stéphane M. Meystre. Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in Health Technology and Informatics*, 264:283, 2019.
- 27 Alexandr Railean. *Improving IoT Device Transparency by Means of Privacy Labels*. PhD thesis, Georg-August-Universität Göttingen, 2022.
- 28 Devin Reich, Ariel Todoki, Rafael Dowsley, Martine De Cock, and Anderson C. A. Nascimento. Privacy-preserving classification of personal text messages with secure multi-party computation. *Advances in Neural Information Processing Systems*, 32, 2019.
- 29 David Sánchez and Montserrat Batet. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163, 2016.
- 30 Björn Schuller and Anton Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2013.
- 31 RJ Senter and Edgar A Smith. Automated readability index. Technical report, Cincinnati Univ OH, 1967.
- 32 Samuel Sousa and Roman Kern. How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review*, pages 1–66, 2022.
- 33 Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. *arXiv preprint arXiv:2011.07018*, 2021.
- 34 Francisco Teixeira, Alberto Abad, Bhiksha Raj, and Isabel Trancoso. Towards end-to-end private automatic speaker recognition. *arXiv preprint arXiv:2206.11750*, 2022.
- 35 Henry Turner, Giulio Lovisotto, and Ivan Martinovic. Speaker anonymization with distribution-preserving x-vector generation for the voiceprivacy challenge 2020. *arXiv preprint arXiv:2010.13457*, 2020.

## 4.6 Uncertain legal interpretation(s) for emerging PETs

Lydia Belkadi (KU Leuven, BE) [lydia.belkadi@kuleuven.be](mailto:lydia.belkadi@kuleuven.be)

Peggy Valcke (KU Leuven, BE) [peggy.valcke@kuleuven.be](mailto:peggy.valcke@kuleuven.be)

License © Creative Commons BY 4.0 International license  
© Lydia Belkadi, Peggy Valcke

Cryptographic-based PETs (e.g., MPC, HE) rely on secret data representations, or modified views of the data, which allow the hiding of the original form of the data. This means that the original form can only be reconstructed by having access to additional sources of data, such as an encryption key, that are held by designated parties. For instance, in HE only authorized parties should have access to the encryption key. In contrast, in MPC several parties hold different, partial representations of the data which, when put together allow for the reconstruction of the original form.



#### 4.6.1 On the legal understanding of information protected by PETs

Within the European Union, data protection rules reflect the scope and aims of a set of fundamental rights, and in particular the rights to privacy and data protection.<sup>17</sup> Indeed, these rights are not absolute, and may suffer limitations so long adequate legal and technical safeguards are adopted.<sup>18</sup> Against this background, EU data protection regimes rely on a broad definition of personal data, while establishing flexibility mechanisms that consider the specific circumstances at stake. This dual objective is reflected in the legal tests data controllers must carry out to determine whether and to what extent data protection rules apply. In legal terms, these tests are defined as the material and territorial scope of the law.

Under EU law, data protection relies on a dynamic conceptual construction between the definitions of personal data, pseudonymized data and anonymized data. For example, Article 2 of the GDPR defines the scope of the law as applying to

- processing of personal data wholly or partly by automated means,
- and processing other than by automated means of personal data which form part of a filing system.

Hence, data protection frameworks only apply to “personal data”, defined as “any information relating to an identified or identifiable natural person” (Article 4(1) of the GDPR). The Article 29 Working Party explains in its Opinion 4/2007 on the Concept of Personal Data that the notion of identification is constructed in a broad way to encompass both direct and indirect identification and identifiability (e.g., name, identifiers, unique combinations of factors).<sup>19</sup>

However, the law also considers the effects of two types of transformations on the legal nature of personal data. On the one hand, anonymous data is explicitly excluded from the material scope of the GDPR. Recital 26 of the GDPR explains that anonymous data is defined negatively, as information which does not fall under the definition of personal data. In other words, anonymous data is information that does not relate to an identified or identifiable individual (i.e., information that is intrinsically anonymous from a data protection perspective) or personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable (i.e., transformed data). The Article 29 Working Party further explains in its Opinion 05/2014 on Anonymisation Techniques that such data transformations should be irreversible. To determine whether a technique amounts to an anonymization processing in the legal sense, the Article 29 Working Party underlines three factors which need to be assessed: singling out, linkability and inference.<sup>20</sup> On the other hand, the GDPR defines pseudonymization as a processing performed in such a manner that personal data cannot be attributed to an individual without the use of additional information (Article 4(5) of the GDPR). In its Opinion 05/2014, the Article 29 Working Party explicitly categorizes encryption with secret keys as a type of pseudonymization processing.<sup>21</sup> The GDPR further requires that additional information should be kept separately and subjected to technical and organizational measures.

PETs have emerged as tools to support controllers in complying with their legal obligations.

<sup>17</sup> The right to private life is enshrined in the Universal Declaration of Human Rights (Article 12), the European Convention on Human Rights (Article 8), and the European Charter of Fundamental Rights (Article 7). Under EU law, a separate right to data protection is also recognized in the European Charter of Fundamental Rights (Article 8).

<sup>18</sup> See the articles referenced above and Article 52 of the European Charter of Fundamental Rights regarding the legal tests that must be performed to restrict the rights to privacy and data protection. In the context of biometric processing, see [1].

<sup>19</sup> Article 29 Working Party, Opinion 4/2007 on the Concept of Personal Data, p. 13.

<sup>20</sup> Article 29 Working Party, Opinion 05/2014 on Anonymisation Techniques p. 7-19.

<sup>21</sup> *Ibid* p. 20.

Nonetheless, there remains substantial uncertainties as to how these technologies interact with and are regulated by data protection rules. This dual aspect of PETs raises important research questions regarding the legal nature of (a) new processing frameworks developed and implemented, and (b) the information protected by PETs.

#### 4.6.2 On the need for specific risk assessments for PETs

In turn, this dual legal nature of PETs also raises important challenges to established legal methodologies to assess data processing risks. In particular, PETs are characterized by their variety and their composite nature. In other words, PETs seek to address specific privacy objectives (e.g., minimization, obfuscation, etc.), and can be used in combination. As a result, the residual risks arising from such assemblages are highly context-dependent. However, there remains a significant gap in interdisciplinary scholarship regarding the analysis and development of tailored risk assessments that would consider the nature and effects of PETs. In particular, there is a significant need to further research venues for interdisciplinary concept-building and suitable flexibility mechanisms.

#### References

- 1 Els Kindt. *Privacy and Data Protection Issues of Biometric Applications. A Comparative Legal Analysis*. Springer, 2013.

## 5 Conclusion

As can be seen from the above findings, the domains of voice, speech, and natural language processing provide a lot of opportunities, but also challenges, when it comes to privacy and data protection. Multiple aspects of privacy have to be considered and incorporated in the design of such processing systems, irrespective of being a smart speaker, voice assistant, or chatbot. The most relevant domains identified in the endeavor of this report are:

- legal considerations (GDPR, EU AI Act, and other applicable laws must be considered),
- human factors (usability and transparency must be addressed, vulnerable groups must be considered),
- technical aspects (voice anonymization techniques and privacy-enhancing technologies in general should be considered whenever possible).

In this report, we provided a first outline of these challenges from an interdisciplinary point of view.

## Participants

- Lydia Belkadi  
KU Leuven, BE
- Zinaida Benenson  
Friedrich-Alexander-Universität –  
Erlangen, DE
- Martine De Cock  
University of Washington –  
Tacoma, US
- Abdullah Elbi  
KU Leuven, BE
- Zekeriya Erkin  
TU Delft, NL
- Natasha Fernandes  
Macquarie University –  
Sydney, AU
- Simone Fischer-Hübner  
Karlstad University, SE
- Ivan Habernal  
TU Darmstadt, DE
- Meiko Jensen  
Karlstad University, SE
- Els Kindt  
KU Leuven, BE
- Dietrich Klakow  
Saarland University –  
Saarbrücken, DE
- Katherine Lee  
Google Brain & Cornell  
University – Ithaca, US
- Anna Leschanowsky  
Fraunhofer IIS – Erlangen, DE
- Pierre Lison  
Norsk Regnesentral – Oslo, NO
- Christina Lohr  
Friedrich-Schiller-Universität –  
Jena, DE
- Emily Mower Provost  
University of Michigan –  
Ann Arbor, US
- Andreas Nautsch  
Avignon Université, FR
- Olga Ohrimenko  
University of Melbourne , AU
- Jo Pierson  
Free University of Brussels, BE
- Laurens Sion  
KU Leuven, BE
- David Stevens  
Gegevensbeschermingsautoriteit –  
Brussels, BE
- Francisco Teixeira  
Instituto Superior Técnico –  
Lisbon, PT
- Natalia Tomashenko  
Avignon Université, FR
- Marc Tommasi  
University of Lille, FR
- Peggy Valcke  
KU Leuven, BE
- Emmanuel Vincent  
Inria – Nancy, FR
- Shomir Wilson  
Pennsylvania State University –  
University Park, US

