Report from Dagstuhl Seminar 22342

Privacy in Speech and Language Technology

Simone Fischer-Hübner[*1], **Dietrich Klakow**[*2], **Peggy Valcke**[*3], **and Emmanuel Vincent**[*4]

1 **Karlstad University, SE.** simone.fischer-huebner@kau.se
2 **Saarland University – Saarbrücken, DE.** dietrich.klakow@lsv.uni-saarland.de
3 **KU Leuven, BE.** peggy.valcke@kuleuven.be
4 **Inria – Nancy, FR.** 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

4 Working Groups

4.1 Case studies and user interaction

Zinaida Benenson (Friedrich-Alexander-Universität – Erlangen, DE) zinaida.benenson@fau.de
Abdullah Elbi (KU Leuven, BE) abdullah.elbi@kuleuven.be
Zekeriya Erkin (TU Delft, NL) z.erkin@tudelft.nl
Natasha Fernandes (Macquarie University – Sydney, AU) natasha.fernandes@mq.edu.au
Simone Fischer-Hübner (Karlstad University, SE) simone.fischer-huebner@kau.se
Ivan Habernal (TU Darmstadt, DE) ivan.habernal@tu-darmstadt.de
Els Kindt (KU Leuven, BE) els.kindt@kuleuven.be
Anna Leschanowsky (Fraunhofer IIS – Erlangen, DE) anna.leschanowsky@iis-extern. fraunhofer.de
Pierre Lison (Norsk Regnesentral – Oslo, NO) plison@nr.no
Christina Lohr (Friedrich-Schiller-Universität – Jena, DE) christina.lohr@uni-jena.de
Emily Mower Provost (University of Michigan – Ann Arbor, US) emilykmp@umich.edu
Jo Pierson (Free University of Brussels, BE) jo.pierson@vub.be
David Stevens (Gegevensbeschermingsautoriteit – Brussels, BE) david.stevens@apd-gba.be
Francisco Teixeira (Instituto Superior Técnico – Lisbon, PT) francisco.s.teixeira@tecnico.ulisboa.pt
Shomir Wilson (Pennsylvania State University – University Park, US) 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.[1]

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).

[1] 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[2] or for public safety or policing purposes (e.g., recognition of aggression in public spaces)[3]. 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.[4] Societies have become data economies with increasing

[2] https://www.vrt.be/vrtnws/nl/2021/09/24/genk/

[3] 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

[4] 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.[5]

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.

[5] See Article 29 Working Party, Opinion on consent, 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).

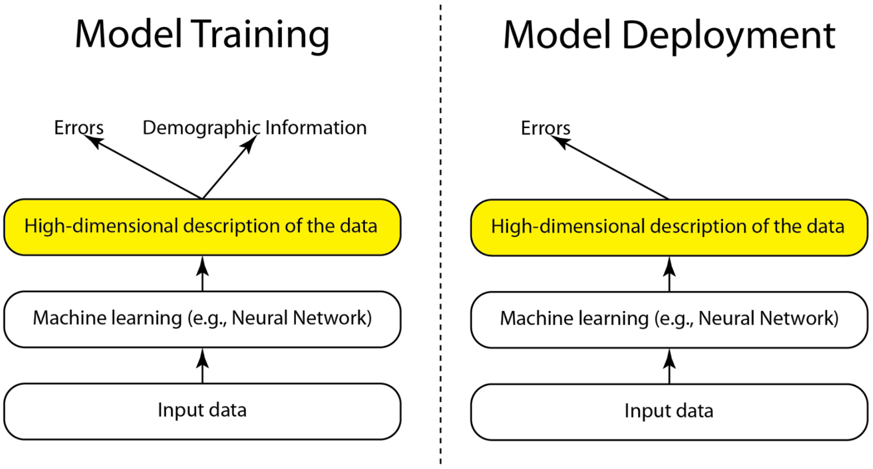

Model Training
Errors
Demographic Information
High-dimensional description of the data
Machine learning (e.g., Neural Network)
Input data
Model Deployment
Errors
High-dimensional description of the data
Machine learning (e.g., Neural Network)
Input data

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?

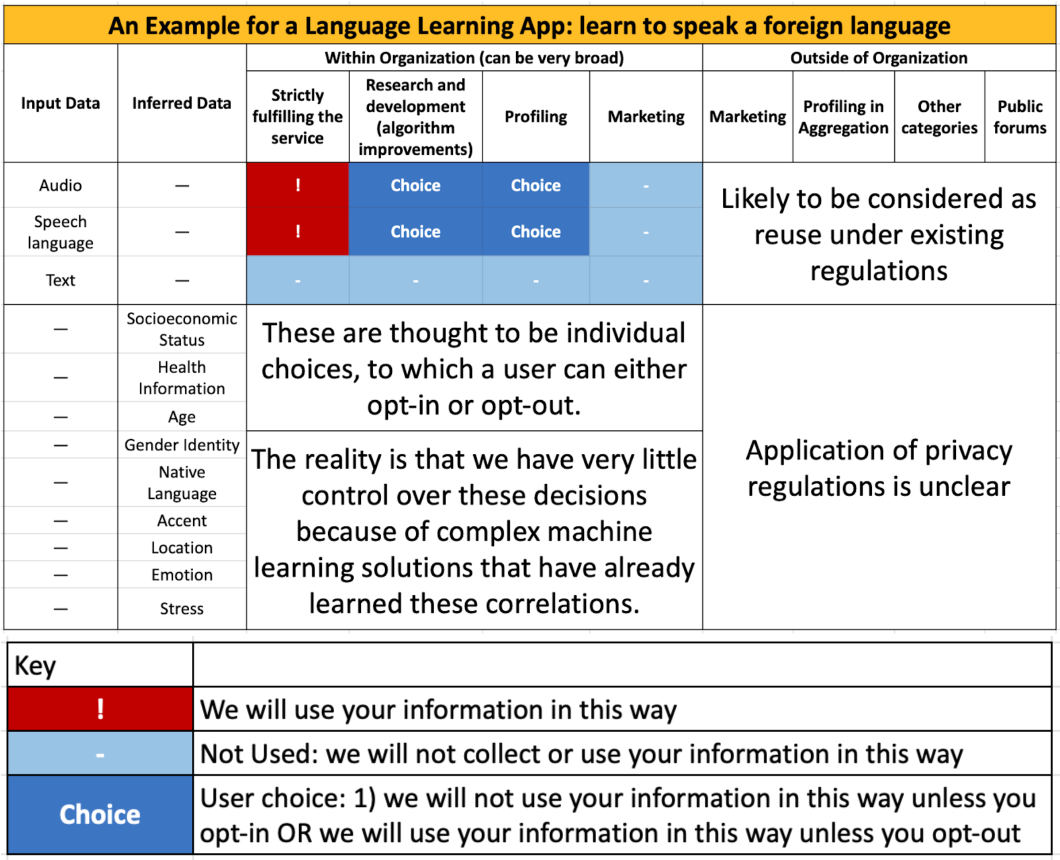
We highlight this challenge in Fig.2 using the example of a language learning app, one that takes in acoustic information and provides feedback to a user to promote the user's language mastery. We assume that the app requires audio information and the ability to extract speech-language information (note 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

		Within Organization (can be very broad)				Outside of Organization			
Input Data	Inferred Data	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								

Key	
!	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 Deatrick, 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
Martine De Cock (University of Washington – Tacoma, US) mdecock@uw.edu
Natasha Fernandes (Macquarie University – Sydney, AU) natasha.fernandes@mq.edu.au
Katherine Lee (Google Brain & Cornell University – Ithaca, US) katherinelee@google.com
Christina Lohr (Friedrich-Schiller-Universität – Jena, DE) christina.lohr@uni-jena.de
Andreas Nautsch (Avignon Université, FR) andreas.nautsch@univ-avignon.fr
Laurens Sion (KU Leuven, BE) laurens.sion@kuleuven.be
Natalia Tomashenko (Avignon Université, FR) natalia.tomashenko@univ-avignon.fr
Marc Tommasi (University of Lille, FR) marc.tommasi@univ-lille.fr
Peggy Valcke (KU Leuven, BE) peggy.valcke@kuleuven.be
Emmanuel Vincent (Inria – Nancy, FR) 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