A Survey on Privacy Issues and Solutions for Voice-controlled Digital Assistants ⋆

Luca Hernández Acosta[1,a], Delphine Reinhardt[a]

[a] *Computer Security and Privacy, University of Göttingen, Göttingen, Germany*

Abstract

With the development and increasing deployment of smart home devices, voice control supports comfortable end user interactions. However, potential end users may refuse to use *Voice-controlled Digital Assistants (VCDAs)* because of privacy concerns. To address these concerns, some manufacturers provide limited privacy-preserving mechanisms for end users; however, these mechanisms are seldom used. We herein provide an analysis of privacy threats resulting from the utilization of VCDAs. We further analyze how existing solutions address these threats considering the principles of the European *General Data Protection Regulation (GDPR)*. Based on our analysis, we propose directions for future research and suggest countermeasures for better privacy protection.

Keywords: Privacy, Voice-controlled Digital Assistants, Smart Speakers, Voice Assistants, Speech Data

⋆The final publication is available at Elsevier via `http://dx.doi.org/10.1016/j.pmcj.2021.101523`

[1] Goldschmidtstr. 7, 37077 Göttingen, Germany, Phone: +49 551 39-172063, Fax: +49 551 39-14403, E-Mail: hernandez@cs.uni-goettingen.de

1. Introduction

VCDAs and their underlying speech recognition technology have improved over time [1]. New application areas for voice-controlled digital assistants have emerged and their overall end user adoption has increased [2]. As a result, voice-controlled user interfaces are now installed on a variety of Internet of Things (IoT) devices, providing intuitive user interactions. They perform various tasks, such as making phone calls and creating shopping lists. And with the advent of smart homes, voice control is also used for more complex tasks. Some of these tasks involve the control of other devices, and the use of these voice-controlled digital assistants in smartphones, in smart speakers, and in connected devices [3]. In the remainder of this paper, we refer to a device equipped with a voice assistant as *VCDA*.

Despite the usefulness of these assistants, their acceptance is hindered by privacy concerns [4]. The first concern is the always listening feature of VCDAs, which gives the impression that the device is constantly active and transmitting recordings to central servers [2, 4]. The second concern is the potential misuse of the personal data collected and processed by manufacturers or third-party developers [5]. In addition to VCDA privacy concerns, some end users have the misconception that their data is processed locally (rather than in the cloud) and not stored, when in fact most providers retain user data indefinitely [5, 6]. Therefore, due to a lack of transparency, there is often a misunderstanding about how end users' data are actually handled.

To protect user privacy and user concerns, different VCDAs allow end users to control the collection and processing of their data. Among them, end users can mute their device via a physical button, adjust privacy settings according to their preferences (e.g., allow access to their name, address, or location), perform audits, and/or manually/automatically delete previous voice recordings. End users, however, often do not know that these options exist or find it too difficult to use them [5, 6]. While they make use of the mute button, some of them do not trust it and prefer to unplug their device [5, 6]. As a result, the available privacy-preserving mechanisms are neither known nor used by most end users and offer only a basic protection. Addressing potential end users' concerns and protecting their privacy is a key aspect to increase the adoption of VCDAs [2, 4].

In this article, our contributions are as follows:

- We consider the different application areas of VCDAs and detail their common architecture including stakeholders in Sec. 2.
- We conduct a privacy and threat analysis showing the nature, location, and extent of such threats in Sec. 3.
- We provide an overview of existing privacy-preserving solutions either partially deployed or proposed in research but not yet adopted. In comparison to previous studies [7, 8], we analyze how these solutions address potential threats and implement the GDPR principles in practice in Sec. 4.
- We finally discuss the limitations of these solutions and identify new research directions in Sec. 5, before concluding in Sec. 6.

2. Foundations of VCDAs

A VCDA is a connected device that allows end users to control it using a voice interface. End users do not need to type their commands, thus opening the doors to a large scope of applications. VCDAs are integrated

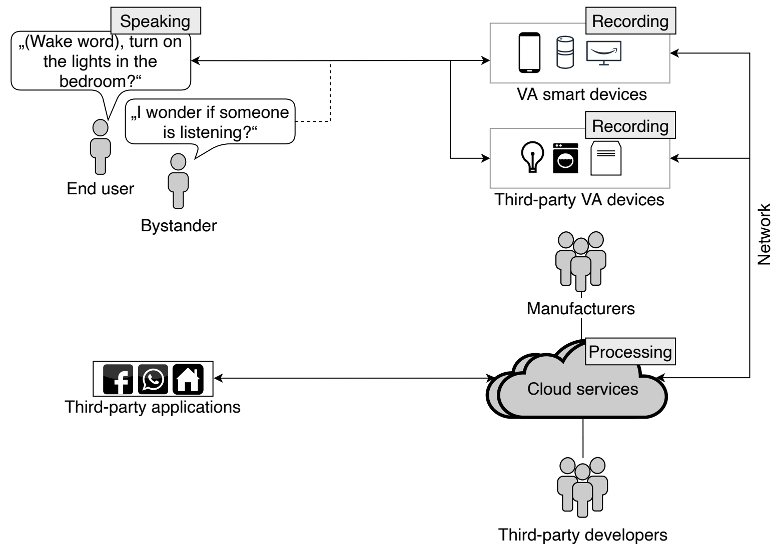

Figure 1: VCDAs ecosystem

in smartphones [1] and smart speakers, which are increasingly being used in private households [2]. Some well-known devices include Amazon Echo Dot, Google Home Mini, and Apple HomePod. These devices are largely used for, e.g., playing music, retrieving information, and controlling household appliances [9]. They are also integrated in vehicles [10], robots [11], televisions, thermostats, and refrigerators [12, 13]. VCDAs are thus implemented in different devices and serve different purposes, while sharing a common goal: Simplifying user interactions. To highlight further similarities, we detail their architecture followed by the stakeholders involved in existing VCDAs.

2.1. Architecture

Whenever end users speak to a VCDA, their voice is first recorded and then forwarded by the VCDA over the Internet to a cloud service [14] as shown in Fig. 1.

On this cloud service, Automated Speech Recognition (ASR) and Natural Language Processing (NLP) processes are executed as basis of the VCDA functionality. ASR has been improved constantly to achieve better performance in converting speech-to-text [15]. In the first step of this conversion, features are extracted from the audio signal using Mel-Frequency Cepstral Coefficients (MFCCs) and then inserted into an acoustic model trained with Hidden Markov Model (HMM). The acoustic model converts the audio signal into the corresponding phonemes, which are then used in a dictionary mapping to form all possible words. Based on large training data sets, a language model checks how high the probabilities are that certain words follow each other to form an intelligible sentence. After the audio signal is converted into a text format, its intent is determined using NLP.

After its interpretation, a response to the initial request is sent to the end user via the VCDA or an action like the playing of music may be performed. Third-party applications can also be triggered by a voice command. Depending on their requirements, the request could be processed on servers or other cloud systems maintained by the third-party developers. The different VCDAs can be categorized according to the following dimensions [16]:

- **Manually activated:** These devices usually require a physical interaction like the pressing of a button for activation e.g. the remote control of a smart TV. After this interaction, the data are forwarded to an online service responsible for data processing and voice recognition.
- **Speech activated:** These devices require a wake word for activation e.g. a smart speaker or a smartphone.
- **Always on:** These devices are active at all time and do not need any interaction for activation e.g. a security camera or a baby monitor.

Note that depending on the environment in which any VCDA is used, bystanders may unknowingly be part of this ecosystem. While the end users intentionally utilize VCDAs, the voices of bystanders may also be recorded.

2.2. Stakeholders

We identify the following stakeholders:

- **End users:** They are individuals who voluntarily use VCDAs. An end user may not necessarily be the purchaser of the VCDA or the holder of the device account but rather one who knowingly uses a VCDA.
- **Bystanders:** In contrast to the voluntary interaction of end users, bystanders' interactions are not voluntary. In other words, bystanders are unaware that their voice is being recorded and processed.
- **Device manufacturers:** They build VCDAs and are responsible for the system architecture and the device hardware. They also process end user requests on servers they operate and maintain. Manufacturers usually have a platform for third-party developers to offer end users new applications that can be downloaded onto these devices.
- **Third-party developers:** They implement new applications and handle end user requests independently of the manufacturer's server.

3. Privacy Analysis

As detailed in Sec. 2.2, different stakeholders interact with the system and these interactions can reveal information about end users and bystanders. To analyze how their privacy can be compromised, we consider each component of the architecture presented in Sec. 2.1. Note that the concept of privacy includes different dimensions [17], ranging from physical to social privacy. Within the scope of this article, we focus on informational privacy following Westin's definition [18], i.e., the right of individuals to control information collected about themselves. As a result, we do not cover invasions to other privacy dimensions that may result from, e.g., unwanted interactions from VCDAs when end users do not want to be disturbed. Moreover, we focus on threats to privacy that are specific to VCDAs. We acknowledge that threats against the communication from and to the VCDAs as well as against the servers from external attackers exist. Nevertheless, these are not specific and can be addressed by applying established security mechanisms [19–21]. In what follows, we therefore consider potential threats coming from internal attackers and analyze them by answering the following questions related to the actors of the VCDA system in order to identify potential attack vectors:

3.1. Who Can Start the Data Collection?

The most common way of starting a VCDA is using a wake word. This word is processed locally, but in some systems it receives support from a wake word verification process in the cloud. Although this process aims at identifying false triggers, smart speakers for example are still falsely activated by similar sounding words [22]. This situation raises privacy concerns for end users who may have private communication that address for e.g. their financial, their emotional, or their health issues recorded without their consent. Unintended voice recordings can also be the result of malicious voice commands [23]. For instance, some companies have used the wake word feature as a marketing gimmick to activate smart speakers when their commercials were on TV [23]. Furthermore, there is also the risk of remote attacks such as the "Dolphin attack" whereby inaudible audio signals are emitted. These signals activate the smart speakers and inject malicious voice commands without the knowledge of the end users [24]. Moreover, malicious applications running on a VCDA can be used to eavesdrop on private conversations and to fish for passwords [25].

3.2. Who Is Collecting, Processing, and Using the Audio Data?

Audio data generated by end users are recorded by the VCDA, then transferred to the servers of the manufacturers and, if necessary, of third-party developers. While end users configuring the VCDA must agree to the terms of use and thus give their consent to use a VCDA, it is not the case for bystanders. Overall, the data collection, processing, and usage remain relatively opaque and it is unclear whether the audio data are accessed by employees for system functionality improvement or other purposes [26]. It is also difficult for end users to obtain information about the data retention period and deletion modalities. Since end users cannot directly delete their recorded data, they also need to trust the manufacturers and third-party developers to actually do it [27].

3.3. Which Other Data Are Affected?

Users' personal information (name, location, phone number, email address, and payment information) and information about other connected devices can also be taken into account when processing users' voice commands [8]. These data can be accessed through applications developed by the manufacturers or by third-party developers and downloaded from the manufacturer's platform. Those applications can be over-privileged and access more data than necessary for their actual purposes [28–30].

3.4. Who Else Can Have Access to These Data?

Depending on the system, end users can have access to their past audio recordings to review and/or delete them. End users can also have access to the recordings of family members, friends, guests at home or even colleagues when VCDAs are deployed in workplaces. Similarly, a request sent to the company to access personal data according to the GDPR Art. 15 can also reveal information about other users [31]. Since voice recognition is not activated by default in the factory settings, individuals other than the end users (e.g., curious or malicious people) can further request the device to disclose sensitive information about end users, like their future appointments, past orders, or personal interests [32].

Potential attacks such as eavesdropping or a man-in-the middle attacks are possible. Such threats are, however, not specific to VCDAs and can be addressed by established security mechanisms [19–21].

3.5. Which Information Can Be Inferred?

Audio data contain a wealth of information about the speakers and their environment. The requested content reveals information about their interests, location, destination, buying behavior, current activities, or mood. Moreover, the uniqueness of their voice allows the linking of their recordings and its characteristics can reveal many insights about the speakers (Tab. 1) [33]. Additional sounds produced by the end users (e.g., coughing and laughing) and background noises (e.g., pets or vehicles) provide further information. Note that the references in Tab. 1 show the feasibility of such inferences, but their original context may differ. Based on these inferred data, a fine-grained portrait of the end users and bystanders can hence be

Table 1: Possible inferences using audio based on [33]

	Main information	Examples of inferred information	References
Characteristics	Gender		[34, 35]
	Age		[36, 37]
	Accent and dialect		[38, 39]
	Body measures	Shapes, height, weight	[40, 41]
	Personality traits	Neuroticism, extraversion, openness	[42, 43]
State	Mood and emotion	Anger, disgust, fear, joy, sadness, surprise	[44, 45]
	Lies		[46]
	Sleepiness		[47, 48]
	Intoxication	Alcohol, drugs	[49, 50]
Health	Physical disorders	Lung, nervous system, voice, communication	[51–53]
	Mental disorders	Depression, schizophrenia, eating disorders	[54, 55]
Social	Interpersonal perception	Persuasiveness, popularity, success	[56, 57]
	Socio-economic status	Social class	[58, 59]
Environment	Presence	Children, pets	[60, 61]
	Activities	TV, cooking, working	[62, 63]

drawn that can be further used by the device manufacturers and third-party developers for profiling and/or for the sending of targeted advertisements.

The respect of the end users' privacy thus mainly depends on the manufacturers and third-party developers incl. associated entities, such as contractors. However, end users can also present a threat to other users' privacy.

4. Privacy-preserving Countermeasures

We present different solutions and discuss their contributions to the implementation of the GDPR principles, and the data subjects' rights.

Table 2: Comparison of existing privacy settings for available devices

Setting	Amazon Alexa	Google Assistant	Apple Siri
Mute device	Yes	Yes	Yes
Notification sounds	Yes	Yes	Yes
View data	Yes	Yes	No
Delete data	Yes	Yes	Yes
Retention period	Indefinite by default, 3 months, 18 months, no retention	Indefinite by default, 3 months, 18 months, 36 months, no retention	Data are stored anonymously, encrypted, and mostly local
Manage skill permissions	Yes	Yes	No
Voice authentication	Yes	Yes	Yes
Opt out	Yes	Yes	Yes

4.1. (Partially) Deployed Solutions

Tab. 2 presents an overview of existing privacy measures for selected platforms. In addition to muting the device and playing a sound when the device is activated to cater for transparency, past recordings can be viewed in the case of Amazon Alexa and Google Assistant. Thus, the *right of access by the data subject* according to GDPR Art. 15 is implemented, albeit only to a very limited extent. According to both privacy policies, voice recordings can also be manually accessed by the manufacturer's employees. However, it is not sufficiently communicated to the end users how exactly their data are processed, so that the principle of *transparency* in GDPR Art. 5 (1)(a) is not fully implemented.

End users of the three selected devices can further delete past recordings and past interactions with their smart home devices. However, both viewing and deleting processes are implemented differently in practice. The primary end user of a device has by default all rights incl. viewing and/ or deleting. In a multi-user environment, secondary end users, or bystanders, have hence no means to protect their privacy from the primary end user or even the manufacturer and third-party developers. Even though most manufacturers allow additional end users to be invited by the primary end user, they can neither view nor delete their data with Alexa and Siri. Only the (primary) end user managing the account can exercise (limited) control over the stored data. Google Assistant is the only platform that allows secondary end users, once invited, to protect their privacy themselves. In this case, primary end users can be prevented from accessing past interactions from others. To the best of our knowledge, Google Assistant also does not further store recordings for people whose voice is not recognized by the system. At least, no recordings are visible to the primary end user. Another issue is the unlimited data retention period set by default for Alexa and Google Assistant, which is against the concept of privacy by default (GDPR Art. 25 (2)) and the principle of *storage limitation* (GDPR Art. 5 (1)(e)). However, end users have the option to automatically delete their recordings after a certain time or not to store them. Siri, on the other hand, stores most data locally and the remaining data are sent anonymized and encrypted to the server(s) [64]. Therefore, the *right to erasure* (GDPR Art. 17) is supported by all manufacturers, even if there exist differences among them.

For Alexa and Google Assistant, primary end users can manage *skills or action permissions* and control the sharing of their personal information, such as address and phone number. Moreover, they can leverage authentication mechanisms, such as *voice match* [65] or *voice profiles* [66]. However, this authentication process can be bypassed by recording and replaying end users' voice [67]. End users can finally opt out of the utilization of their data for additional purposes, such as system improvements.

4.2. Suggested Solutions

We explore additional countermeasures to protect end user's privacy that have not been deployed yet. We consider these solutions based on the threats identified in Sec. 3 and apply them to architecture depicted in Fig. 2. They further serve as basis for identifying future research directions (Sec. 5).

4.2.1. Voice and Speech Protection

The information in a speech signal consists of voice timbre, prosody, and speech content [68]. These components allow inferring different users' characteristics (Tab. 1). By altering them, the speaker's identity and personal attributes can be protected using an intermediary application such as *VoiceMask* [69]. This application first converts and anonymizes end users' voice using a frequency warping process based on *Vocal Tract Length Normalization (VTLN)* [70]. By stretching and compressing the spectrum in relation to the frequency axis, a new voice is generated that transmits the same speech content [69]. After the modification of the speaker's voice, keywords are substituted by identifying sensitive words predefined by the end users and replaced by safewords. While VoiceMask keeps track of replaced words, its cloud service does not notice these word replacements and end users can thus control which words are shared with it.

After the voice recording, the recorded speech command can be sanitized by VoiceMask using voice conversion and keyword substitution (see A in Fig. 2). The sanitized command is then sent to the manufacturer's cloud for processing and forwarded to the corresponding application if necessary. As a result, potential eavesdroppers, attackers, as well as manufacturers and

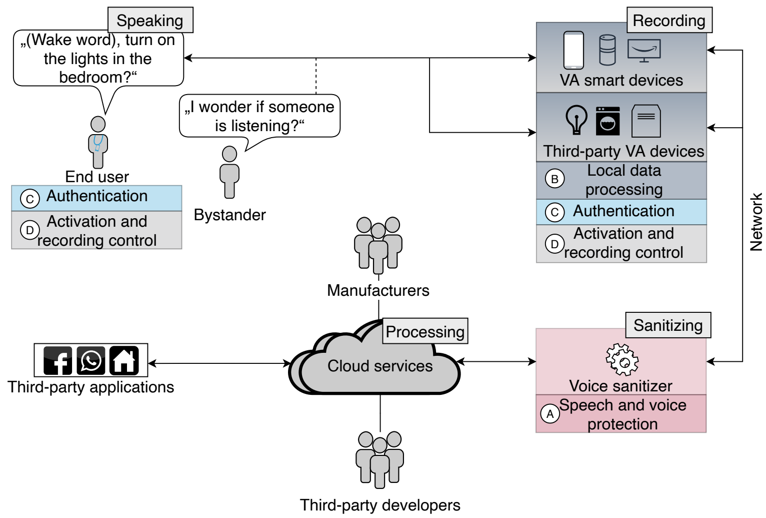

Figure 2: Privacy-preserving countermeasures applied to the current architecture

third-party developers are unable to identify end users and infer personal characteristics based on their voice. Voice conversion reduces features that would allow identity determination as well as the derivation of other personal information such as gender, age or health aspects. As a result, the processing purpose is limited a priori and only the command itself is transmitted, which is in-line with the GDPR Art. 5 (1)(b) principle of *purpose limitation*. Keyword substitution, on the other hand, allows sensitive information about the user to be protected in advance and not sent to the provider in accordance with the GDPR Art. 5 (1)(c) principle of *data minimization*.

4.2.2. Local Data Processing

Another solution to prevent third-parties and/or manufacturers from obtaining end users' information is to locally process the data on the device (see B in Fig. 2). To this end, Coucke et al. have developed a system, named *Snips* that allows for offline processing of voice commands [71]. They have shown that it is feasible to implement their solution on resource-limited IoT devices, such as a Raspberry Pi 3. In order to ensure state-of-the-art performance while simultaneously protecting privacy, they rely on crowdsourcing data and semi-supervised machine learning instead of the users' own data, thus creating a data generation pipeline [71]. To do so, a pre-trained model tailored to specific tasks is required and influences the functionality of Snips. Depending on the nature of the end user's commands, data can still be sent to other cloud services, thus limiting the provided privacy protection. By using Snips, the field of application of VCDAs is hence limited as compared to their original version due to the specific task models. *Rhasspy*, an alternative to Snips, is an offline, open-source and privacy-friendly voice assistant service platform which is currently under development [72]. It targets a mechanism for setting up voice control applications.

By processing data locally, most user data are no longer sent to providers for processing and storage. As a result, the GDPR Art. 5 (1)(c) principles of *data minimization* and Art. 5 (1)(e) *storage limitation* are better applied as compared to current solutions with whom data are stored indefinitely.

4.2.3. Authentication

Different methods have been developed to prevent attackers from obtaining end users' recordings, or from gaining access to end users' device resource, or from controlling their device function. The first method requires the end user to wear an additional device that measures body vibrations when speaking [73]. These body vibration data in combination with the end user command are used to authenticate the end user. Carrying additional hardware can however be inconvenient for end users and can potentially reveal further information about them.

To overcome this limitation, *VoiceGesture* [74] has been developed. It relies on an analysis of vocal articulatory gestures involving the movement of different articulators like upper lip, lower lip or jaw according to the pronounced phonemes based on the Doppler effect [74]. While this technique is promising and offers reliable authentication rates, it has only been tested with smartphones and two different positions near the end user's mouth, thus limiting its current applicability.

Another solution exploits the difference between the voice from a human speaker and that from a loudspeaker [75]. The drawback of this method is that other magnetic fields close to the end user can cause interference leading to errors [75]. Also, the software is limited by the distance of the loudspeaker to the VCDA: The longer the distance, the worse the performance [75].

As a result, the integration of the proposed methods in the architecture would require either the end user to carry an additional device or the VCDA to distinguish legitimate end users from replayed recordings (Ⓒ in Fig. 2). By implementing these solutions, it will become more difficult for persons without authorization to gain access to audio data of legitimate users. Although these mechanisms address threats originating from users located in proximity to the device, they do not offer protection against other threats coming from data processing entities, such as manufacturers or third-party developers. As a result, they only partly address GDPR Art. 5(1)(f) *integrity and confidentiality.*

4.2.4. Activation and Recording Control

Instead of turning off or even unplugging the device, Mhaidli et al. presented a prototype that is based on volume and gaze to control when a VCDA should start to listen [76]. When using it, the VCDA will not send and process unintended audio in the cloud. Therefore, the proposed solution also relies on analyzing the speaker's gaze direction using a camera to complement the volume-based control. Consequently, this solution could be integrated in the VCDA (see D in Fig. 2) and would contribute to better control when recordings are made. As fewer unintentional activations occur, less data is sent to cloud services, so data is only processed when it is truly intended, contributing to the GDPR Art. 5 (1)(c) principle of *data minimization*. A limitation for this method is that it requires the deployment of a camera that would also collect information about the end users.

Another approach proposed by Cheng et al. [77] is an acoustic tagging system that provides advanced control over what audio data should be processed in the cloud [77]. The system requires an additional tagging device that also listens for the wake word and emits an acoustic tag that is detected by the VCDA, which will in turn potentially delete the recording. Audio jamming is also another technique suggested by Gao et al., [78], to prevent involuntary audio recording and maintain user privacy.

In summary, most of the proposed solutions focus on reducing end users' information which is transferred to manufacturers and third-party developers. These solutions apply different techniques that mask voice, replace words, and locally process data. Other solutions provide more control over their audio recordings to end users, and limit the risk for bystander audio recording.

5. Discussion

In this section, we first discuss the limitations of existing solutions and propose new research directions, before making comments on the role of privacy in the acceptance of VCDAs and a comparison with other technologies.

5.1. Limitations of Existing Solutions

The current available solutions and the ones proposed in research offer only a limited privacy protection to both end users and bystanders. Unexpected activations of VCDAs can lead to non-consensual recording of private conversations. While manufacturers are continuously improving the processing of wake words, unintentional activations are still possible. As presented in Sec. 4.2.4, the integration of a camera to detect the direction of the end users' gaze has been proposed to reduce such false positives; however, such integration may also present additional end user privacy threats.

While the options for the mitigation of privacy threats vary among manufacturers, the solutions regulating the access to the recordings and associated personal data are unsatisfactory. Some do not allow end users to audit or delete past recordings. And even if these options are available, the protection against unauthorized access by honest-but-curious or active attackers located close to these devices is insufficient. For example, although Google Assistant does not allow an unrecognized voice access to an end user's personal information after the *voice match* option has been set up, simply recording and playing back the wake word of the legitimate end user's voice can bypass this privacy control. Moreover, Google Assistant's voice match feature is not set by default but requires end users to activate this option. Thus, end users of Google Assistant unaware of the voice match option to mitigate such privacy threats are exposed to an even higher risk.

The solutions discussed in Sec. 4.2.3 can reduce such replay attacks, but they still need to collect information about the end users. Also, each of these solutions require an additional device and the practicability of their deployment in real-world scenarios is still unclear. Moreover, these solutions do not permit bystanders in a multi-user environment to protect their privacy from the holder of the main account except when separate accounts exist. In this case, bystanders have to register with their own voice assistant account (for e.g. with Alexa or Google Assistant). However, they will still depend on the primary user to set the appropriate settings.

The solutions presented in Sec. 4.2.2 and Sec. 4.2.1 involve the local processing of voice to prevent their disclosure to device managers and the disguising of voice to conceal them from third-party developers respectively. Also, these solutions can impact the performance of the command recognition, thus potentially limiting the usability and potential adoption of future users. For example, the replacement of words by others according to the end users' preferences requires end users to train their device with these words beforehand—an additional drawback in terms of usability.

5.2. Future Research Directions

Already implemented solutions offer only limited possibilities to protect the privacy of end users as well as the rights of bystanders to exercise their data protection rights. In the specific case of bystanders, data can be collected about them without their knowledge and consent that is against Art. 5 (1)(a) and Art. 6 (1)(a). In absence of knowledge of such data collection, it is therefore difficult for bystanders to enforce their rights defined in Art. 12 to 23. For example, it includes the right of access by the data subject (Art. 15) or the right to erasure (Art. 17) granted by the GDPR. The development of solutions to allow bystanders to exercise these rights is therefore necessary. Designing such solutions is however challenging due to the potential dynamics of such brief interactions. Moreover, solutions that require additional privacy control devices for bystanders or efforts from the bystanders should be reduced to the minimum to make them more acceptable by potential users.

We argue that end users should be able to exercise their rights in a multi-user environment [79]. Especially rights to data access (Art. 15) and data erasure (Art. 17) which is often available to only the primary end user maintaining the main account and not other users. Each user could set up their own account incl. their voice profile, so that their interactions can be linked together and only be accessible by them. On the other hand, non-registered voices should not be stored to protect bystanders' privacy, at least from other end users who have set up an account on the VCDA. However, (online) voice recognition still requires the collection and processing of audio recordings. As a result, manufacturers and third-party developers would still have access to these data. An offline and local solution would solve it, assuming that it is protected against attackers located nearby.

Moreover, a wealth of data incl. audio data about the end users are linked to their accounts. As a result, we propose to reduce the amount of data stored by the manufacturers to the minimum by design and give a better control to end users and bystanders over the data collection. To this end, a dedicated application installed on their phone will serve as a broker between them and the VCDA to manage their preferences. It will generate a unique but not reproducible token that identifies the person with the VCDA. Depending on the person's preferences, a voice profile will then be generated on the smartphone and shared with the VCDA to give explicit consent to data collection and processing. Without a shared voice profile, the VCDA would not be operable, so that the privacy of the concerned people would be protected. By providing the identification token, users having given their consent for collecting and processing will be able to access the history of their requests stored offline on the VCDA using their smartphone. Note that we will respect both privacy by design and privacy by default principles in the solution to be designed.

While the settings of VCDAs should be the most privacy-preserving by default according to the GDPR (Art. 25), Lau et al. [5, 80] and Malkin et al. [6] have shown that accessing and configuring them remain a complex task for most end users. Therefore, we recommend that settings such as saving the history of voice commands or using voice recordings to improve the system are not enabled by default to comply with Art. 25 (2). In addition, we encourage the development of new prototypes that do not collect user data *by design* (Art. 25 (1)).

Finally, current solutions and studies conducted in this area primarily focus on the deployment of VCDAs in home scenarios. However, such devices can also be deployed in workplaces. Since the concept of privacy is context-dependent, additional efforts should be conducted, not only to better understand the associated privacy concerns and their impact on the adoption of these devices, but also by researching solutions to address them.

5.3. Privacy and Acceptance

When using VCDAs, end users can gain benefits (i.e., easier interactions, new possible controls), while simultaneously putting their privacy at risk (i.e., disclosing information about themselves to the device). There exists therefore a trade-off between their utility and privacy risks [81]. Such trade-off, however, depends on both the concerned individuals and the application context. For example, using voice commands with various devices, such as turning lights on and off, can be a significant benefit for people with physical disabilities and can contribute to an improvement in their quality of life. Nevertheless, it is worth considering whether the amount of information collected and inferred by such devices are commensurate with their benefits. Even if end users are aware of potential risks, they can still decide to use the device—an illustration of the known *privacy paradox.* The reasons behind their choice can be utilitarian, hedonic, symbolic, or due to social presence, or social attraction [82]. The privacy paradox is however not specific to VCDAs; it is also observed with e.g. robots where users are willing to trade their privacy for given benefits (e.g. [83]). Despite privacy concerns, end users can however only accept the system as it is. They have only limited means to verify that providers apply appropriate security measures and must therefore trust them, as the provided protection is not transparent for them. Nevertheless, providers should remain the ones responsible for secure processing and privacy protection. Conversely, end users should not be blamed for not using privacy protection methods because these are too complex. Instead, it should be our objective as a community to design usable solutions that provide more privacy protection, control and transparency.

5.4. Surveillance Potential and Differences with Other Technologies

VCDAs can record the voice of people without their consent in different scenarios (see Sec. 3). Hence, they can be misused and can even serve as surveillance technology [84]. However, it is interesting to observe the differences between VCDAs and other surveillance technologies, such as CCTV cameras. For example, the use of cameras in public places is strictly regulated [85], while it is not the case for VCDAs. Similarly, there are currently no requirements to report the use of VCDAs or microphones in general, unlike cameras. The use of microphones in public places because of the threats to privacy (see Sec. 3); however, raises similar concerns as cameras but is handled differently in the legislation. As such, political discussions and harmonization efforts should therefore be conducted in this direction.

6. Conclusions

In this article, we have first described the architecture and the stakeholders that interact with each other when using VCDAs. We have next highlighted which data are collected about end users, how they are processed, and who has access to these data. We have then analyzed which privacy-preserving measures are already implemented in three selected cases, namely Alexa, Siri, and Google Assistant. In our analysis, we have further discussed how and to what extent these existing measures allow end users to exercise their rights according to the GDPR. While privacy settings and functions, such as viewing and deleting interactions, with a VCDA are available, they are however limited, not set by default, unknown to the end users and often difficult to use by them. As a result, end users only rarely use them. In addition, past private interactions can be protected from access by other users by leveraging voice recognition as authentication method. This is especially helpful in multi-user environments, where each interaction can be linked to the respective user's account (instead of the one of the primary end user). However, this feature is not supported by all manufacturers yet.

To protect users' privacy, different mechanisms presented in this article have been proposed. They include solutions to (1) obfuscate a user's voice and remove sensitive voice characteristics, (2) substitute sensitive user-defined words, (3) locally process the commands (instead of on external servers), (4) authenticate users, and (5) control when the audio recording starts and what data should be processed online. Their integration into end products would further contribute to the application of the GDPR principles detailed in Art. 5 (1)(b), (1)(c), and (1)(e) in practice.

Consequently, few approaches to better protect both end users and bystanders have been integrated in commercial off-the-shelf products and different solutions have been proposed in research, some of them still suffering from limitations about their practicability discussed in this article. As a result, more efforts are required from both our research community as well as companies to provide and implement solutions that offer better privacy protection and comply with the GDPR, while being usable. These efforts are important and will allow to answer additional questions that arise from the increasing number of VCDAs deployed in our environment at home, at workplaces, or shops. Examples of such questions are: (1) how can the processing of voice data be made transparent in absence of visual interfaces, (2) how can we improve the usability of the privacy settings by reducing the number of user interactions to the minimum while still supporting comprehension and providing control, (3) how can we better inform bystanders about the deployment of VCDAs and allow them to give explicit concern and better control the processing, viewing, and deleting of their data, and (4) how can the different individual privacy conceptions be taken into account in multi-user environments. By addressing these questions, the ultimate goal is to continue to benefit from the advantages offered by VCDAs, while protecting users' and bystanders privacy.

Acknowledgments

We thank the anonymous reviewers for their feedback that have contributed to improve the quality of this article.

References

[1] M. B. Hoy, Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants, Medical Reference Services Quarterly (2018).

[2] S. Han, H. Yang, Understanding Adoption of Intelligent Personal Assistants, Industrial Management & Data Systems (2018).

[3] M. Porcheron, J. E. Fischer, S. Reeves, S. Sharples, Voice Interfaces in Everyday Life, in: Proc. of the 18th Conf. on Human Factors in Computing Systems (CHI), 2018.

[4] P. Kowalczuk, Consumer Acceptance of Smart Speakers: A Mixed Methods Approach, Journal of Research in Interactive Marketing (2018).

[5] J. Lau, B. Zimmerman, F. Schaub, Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers, Proc. of the ACM on Hum.-Comp. Interact. (2018).

[6] N. Malkin, J. Deatrick, A. Tong, P. Wijesekera, S. Egelman, D. Wagner, Privacy Attitudes of Smart Speaker Users, Proc. on Privacy Enhancing Technologies (PoPETs) (2019).

[7] N. Abdi, K. M. Ramokapane, J. M. Such, More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assis-

tants, in: Proc. of the 15th Symposium on Usable Privacy and Security (SOUPS), 2019.

[8] J. S. Edu, J. M. Such, G. Suarez-Tangil, Smart Home Personal Assistants: A Security and Privacy Review, ACM Comput. Surv. (2020).

[9] F. Bentley, C. Luvogt, M. Silverman, R. Wirasinghe, B. White, D. Lottridge, Understanding the Long-Term Use of Smart Speaker Assistants, Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (2018).

[10] M. Braun, A. Mainz, R. Chadowitz, B. Pfleging, F. Alt, At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces, in: Proc. of the 46th Conf. on Human Factors in Computing Systems (CHI), 2019.

[11] S. Poirier, F. Routhier, A. Campeau-Lecours, Voice Control Interface Prototype for Assistive Robots for People Living with Upper Limb Disabilities, in: Proc. of the 16th IEEE Int. Conf. on Rehabilitation Robotics (ICORR), 2019.

[12] Honeywell, WI-FI Smart Thermostat With Voice Control, 2017. [Online]. https://www.honeywellstore.com/store/products/wi-fi-smart-thermostat-with-voice-rth9590wf1003u.htm, accessed in 2021-08-05.

[13] Samsung, Family Hub 2.0: Kühlen Einen Schritt Weitergedacht, 2017. [Online]. https://news.samsung.com/de/family-hub-2-0-kuhlen-einen-schritt-weitergedacht, accessed in 2021-08-05.

[14] H. Chung, J. Park, S. Lee, Digital Forensic Approaches for Amazon Alexa Ecosystem, Digital Investigation (2017).

[15] D. Yu, L. Deng, Automatic Speech Recognition., 2016.

[16] S. Gray, Always On: Privacy Implications of Microphone-Enabled Devices, in: Proc. of the 2nd Future of Privacy Forum (FPF), 2016.

[17] H. Leino-Kilpi, M. Välimäki, T. Dassen, M. Gasull, C. Lemonidou, A. Scott, M. Arndt, Privacy: A Review Literature, Int. Journal of Nursing Studies (2001).

[18] A. F. Westin, Privacy and Freedom, Washington and Lee Law Review (1968).

[19] Z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, S. Shieh, IoT Security: Ongoing Challenges and Research Opportunities, in: Proc. of the 7th IEEE Int. Conf. on Service-oriented Computing and Applications (SOCA), 2014.

[20] M. A. Khan, K. Salah, IoT Security: Review, Blockchain Solutions, and Open Challenges, Future Generation Computer Systems (2018).

[21] V. Adat, B. B. Gupta, Security in Internet of Things: Issues, Challenges, Taxonomy, and Architecture, Telecommunication Systems (2018).

[22] L. Schönherr, M. Golla, T. Eisenhofer, J. Wiele, D. Kolossa, T. Holz, Unacceptable, Where Is My Privacy? Exploring Accidental Triggers of Smart Speakers, arXiv preprint arXiv:2008.00508 (2020).

[23] H. Chung, M. Iorga, J. Voas, S. Lee, Alexa, Can I Trust You?, Computer (2017).

[24] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, W. Xu, Dolphinattack: Inaudible Voice Commands, in: Proc. of the 24th ACM SIGSAC Conf. on Computer and Communications Security (CCS), 2017.

[25] F. Bräunlein, L. Frerichs, Smart Spies: Alexa and Google Home Expose Users to Vishing and Eavesdropping, 2019. [Online]. `https://www.srlabs.de/bites/smart-spies`, accessed in 2021-08-05.

[26] BBC, Smart Speaker Recordings Reviewed by Humans, 2019. [Online]. `https://www.bbc.com/news/technology-47893082`, accessed in 2021-08-05.

[27] T. Mather, S. Kumaraswamy, S. Latif, Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance, "O'Reilly Media, Inc.", 2009.

[28] A. Natatsuka, R. Iijima, T. Watanabe, M. Akiyama, T. Sakai, T. Mori, Poster: A First Look At the Privacy Risks of Voice Assistant Apps, in: Proc. of the 12th ACM SIGSAC Conf. on Computer and Communications Security (CCS), 2019.

[29] D. Su, J. Liu, S. Zhu, X. Wang, W. Wang, "Are You Home Alone?" "Yes" Disclosing Security and Privacy Vulnerabilities in Alexa Skills, arXiv preprint arXiv:2010.10788 (2020).

[30] E. Fernandes, J. Jung, A. Prakash, Security Analysis of Emerging Smart Home Applications, in: Proc. of the 37th IEEE Symposium on Security and Privacy (S&P), 2016.

[31] M. Day, G. Turner, N. Drozdiak, Alexa, Who Has Access to My Data? Amazon Reveals Private Voice Data Files, 2018. [Online]. https://www.heise.de/downloads/18/2/5/6/5/3/9/6/ct.0119.016-018_engl.pdf, accessed in 2021-08-05.

[32] N. AlOtaibi, F. Lombardi, Privacy and Security Evaluation of Amazon Echo Voice Assistant, in: Proc. of the 1st IEEE Int. Conf. of Women in Data Science at Taif University (WiDSTaif), 2021.

[33] J. L. Kröger, O. H.-M. Lutz, P. Raschke, Privacy Implications of Voice and Speech Analysis–Information Disclosure by Inference, in: Proc. of the 14th IFIP Int. Summer School on Privacy and Identity Management (IFIP SC), 2019.

[34] E. Mendoza, N. Valencia, J. Muñoz, H. Trujillo, Differences in Voice Quality Between Men and Women: Use of the Long-Term Average Spectrum (LTAS), Journal of Voice (1996).

[35] S. H. Kabil, H. Muckenhirn, M. Magimai-Doss, On Learning to Identify Genders From Raw Speech Signal Using CNNS., in: Proc. of the 19th Interspeech, 2018.

[36] P. H. Ptacek, E. K. Sander, Age Recognition from Voice, Journal of Speech and Hearing Research (1966).

[37] S. O. Sadjadi, S. Ganapathy, J. W. Pelecanos, Speaker Age Estimation on Conversational Telephone Speech using Senone Posterior Based I-Vectors, in: Proc. of the 41st IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2016.

[38] F. Biadsy, Automatic Dialect and Accent Recognition and its Application to Speech Recognition, Ph.D. thesis, Columbia University, 2011.

[39] M. A. Zissman, T. P. Gleason, D. Rekart, B. L. Losiewicz, Automatic Dialect Identification of Extemporaneous Conversational, Latin American Spanish Speech, in: Proc. of the 21st IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 1996.

[40] R. M. Krauss, R. Freyberg, E. Morsella, Inferring Speakers' Physical Attributes from their Voices, Journal of Experimental Social Psychology (2002).

[41] I. Mporas, T. Ganchev, Estimation of Unknown Speaker's Height from Speech, Int. Journal of Speech Technology (2009).

[42] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, et al., A Survey on Perceived Speaker Traits: Personality, Likability, Pathology, and the First Challenge, Computer Speech & Language (2015).

[43] T. Polzehl, Personality in Speech., 2016.

[44] B. Schuller, G. Rigoll, M. Lang, Hidden Markov Model-Based Speech Emotion Recognition, in: Proc. of the 28th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2003.

[45] T. L. Nwe, S. W. Foo, L. C. De Silva, Speech Emotion Recognition using Hidden Markov Models, Speech Communication (2003).

[46] G. Mendels, S. I. Levitan, K.-Z. Lee, J. Hirschberg, Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection, in: Proc. of the 18th Interspeech, 2017.

[47] H. P. Greeley, E. Friets, J. P. Wilson, S. Raghavan, J. Picone, J. Berg, Detecting Fatigue from Voice using Speech Recognition, in: Proc. of the 6th IEEE Int. Symposium on Signal Processing and Information Technology (ISSPIT), 2006.

[48] J. Krajewski, B. Kröger, Using Prosodic and Spectral Characteristics for Sleepiness Detection, in: Proc. of the 8th Annual Conf. of the Int. Speech Communication Association (ISCA), 2007.

[49] S. Ultes, A. Schmitt, W. Minker, Attention, Sobriety Checkpoint! Can Humans Determine by Means of Voice, if Someone is Drunk... and can Automatic Classifiers Compete?, in: Proc. of the 12th Annual Conf. of the Int. Speech Communication Association (ISCA), 2011.

[50] G. Bedi, G. A. Cecchi, D. F. Slezak, F. Carrillo, M. Sigman, H. De Wit, A Window into the Intoxicated Mind? Speech as an Index of Psychoactive Drug Effects, Neuropsychopharmacology (2014).

[51] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, S. N. Patel, Accurate and Privacy Preserving Cough Sensing using a Low-Cost Microphone, in: Proc. of the 13th Int. Conf. on Ubiquitous Computing (UbiComp), 2011.

[52] J. Rusz, R. Cmejla, H. Ruzickova, E. Ruzicka, Quantitative Acoustic Measurements for Characterization of Speech and Voice Disorders in Early Untreated Parkinson's Disease, The Journal of the Acoustical Society of America (2011).

[53] F. Kazinczi, K. Mészáros, K. Vicsi, Automatic Detection of Voice Dis-

orders, in: Proc. of the 3rd Int. Conf. on Statistical Language and Speech Processing (SLSP), 2015.

[54] D. M. Low, K. H. Bentley, S. S. Ghosh, Automated Assessment of Psychiatric Disorders using Speech: A Systematic Review, Laryngoscope Investigative Otolaryngology (2020).

[55] C. M. Corcoran, F. Carrillo, D. Fernández-Slezak, G. Bedi, C. Klim, D. C. Javitt, C. E. Bearden, G. A. Cecchi, Prediction of Psychosis Across Protocols and Risk Cohorts Using Automated Language Analysis, World Psychiatry (2018).

[56] C. A. Klofstad, R. C. Anderson, S. Peters, Sounds like a Winner: Voice Pitch Influences Perception of Leadership Capacity in Both Men and Women, Proc. of the Royal Society B: Biological Sciences (2012).

[57] N. Miller, G. Maruyama, R. J. Beaber, K. Valone, Speed of Speech and Persuasion, Journal of Personality and Social Psychology (1976).

[58] M. L. Rowe, Child-Directed Speech: Relation to Socioeconomic Status, Knowledge of Child Development and Child Vocabulary Skill, Journal of Child Language (2008).

[59] B. Bernstein, Language and Social Class, The British Journal of Sociology (1960).

[60] A. Mesaros, T. Heittola, T. Virtanen, TUT Database for Acoustic Scene Classification and Sound Event Detection, in: Proc. of the 24th European Signal Processing Conf. (EUSIPCO), 2016.

[61] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, T. Virtanen, DCASE Challenge Setup: Tasks, Datasets and

Baseline System, in: Proc. of the 2nd Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[62] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, M. D. Plumbley, Detection and Classification of Acoustic Scenes and Events, IEEE Transactions on Multimedia (2015).

[63] J. M. Sim, Y. Lee, O. Kwon, Acoustic Sensor based Recognition of Human Activity in Everyday Life for Smart Home Services, Int. Journal of Distributed Sensor Networks (2015).

[64] Apple, We're Committed to Protecting Your Data., 2021. [Online]. `https://www.apple.com/privacy/features/`, accessed in 2021-08-05.

[65] Google, Link Your Voice to Your Devices with Voice Match, 2021. [Online]. `https://support.google.com/assistant/answer/9071681`, accessed in 2021-08-05.

[66] Amazon, What Are Alexa Voice Profiles?, 2021. [Online]. `https://www.amazon.com/gp/help/customer/display.html?nodeId=202199440`, accessed in 2021-08-05.

[67] D. Anniappa, Y. Kim, Security and Privacy Issues With Virtual Private Voice Assistants, in: Proc. of the 11th IEEE Annual Computing and Communication Workshop and Conf. (CCWC), 2021.

[68] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li, Spoofing and Countermeasures for Speaker Verification: A Survey, Speech Communication (2015).

[69] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X.-Y. Li, Y. Wang, Y. Deng, Voicemask: Anonymize and Sanitize Voice Input on Mobile Devices, arXiv preprint arXiv:1711.11460 (2017).

[70] J. Cohen, T. Kamm, A. G. Andreou, Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability, The Journal of the Acoustical Society of America (1995).

[71] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, et al., Snips Voice Platform: An Embedded Spoken Language Understanding System for Private-by-Design Voice Interfaces, arXiv preprint arXiv:1805.10190 (2018).

[72] Rhasspy, Rhasspy Voice Assistant, 2019. [Online]. `https://rhasspy.readthedocs.io/en/latest/`, accessed in 2021-08-05.

[73] H. Feng, K. Fawaz, K. G. Shin, Continuous Authentication for Voice Assistants, in: Proc. of the 23rd Annual Int. Conf. on Mobile Computing and Networking (MobiCom), 2017.

[74] L. Zhang, S. Tan, J. Yang, Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication, in: Proc. of the 24th ACM SIGSAC Conf. on Computer and Communications Security (CCS), 2017.

[75] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, A. Mohaisen, You Can Hear but You Cannot Steal: Defending Against Voice Impersonation Attacks on Smartphones, in: Proc. of the 37th IEEE Int. Conf. on Distributed Computing Systems (ICDCS), 2017.

[76] A. Mhaidli, M. K. Venkatesh, Y. Zou, F. Schaub, Listen Only When Spoken To: Interpersonal Communication Cues as Smart Speaker Privacy Controls, Proc. on Privacy Enhancing Technologies (PoPETs) (2020).

[77] P. Cheng, I. E. Bagci, J. Yan, U. Roedig, Smart Speaker Privacy Control-Acoustic Tagging For Personal Voice Assistants, in: Proc. of the 40th IEEE Security and Privacy Workshops (SPW), 2019.

[78] C. Gao, V. Chandrasekaran, K. Fawaz, S. Banerjee, Traversing the Quagmire That Is Privacy in Your Smart Home, in: Proc. of the 2nd Workshop on IoT Security and Privacy (IoT S&P), 2018.

[79] N. Meng, D. Keküllüoğlu, K. Vaniea, Owning and Sharing: Privacy Perceptions of Smart Speaker Users, Proc. of the ACM on Human-Computer Interaction (HCI) (2021).

[80] J. Lau, B. Zimmerman, F. Schaub, "Alexa, Stop Recording": Mismatches between Smart Speaker Privacy Controls and User Needs, in: Posters at the 14th Symposium on Usable Privacy and Security (SOUPS), 2018.

[81] G. Chalhoub, I. Flechais, "Alexa, Are You Spying On Me?": Exploring the Effect of User Experience on the Security and Privacy of Smart Speaker Users, in: Proc. of the 22nd Int. Conf. on Human-Computer Interaction (HCI), 2020.

[82] C. Lutz, G. Newlands, Privacy and Smart Speakers: A Multi-Dimensional Approach, The Information Society (2021).

[83] C. Lutz, A. Tamò-Larrieux, The Robot Privacy Paradox: Understanding How Privacy Concerns Shape Intentions to Use Social Robots, Human-Machine Communication (2020).

[84] E. West, Amazon: Surveillance as a Service, Surveillance & Society (2019).

[85] A. Šidlauskas, Video Surveillance and the GDPR, in: Proc. of the 7th Social Transformations in Contemporary Society (STICS), 2019.