What is Sensitive About (Sensitive) Data? Characterizing Sensitivity and Intimacy with Google Assistant Users

Gómez Ortega, Alejandra; Bourgeois, Jacky; Kortuem, Gerd

DOI
10.1145/3544548.3581164

Publication date
2023

Document Version
Final published version

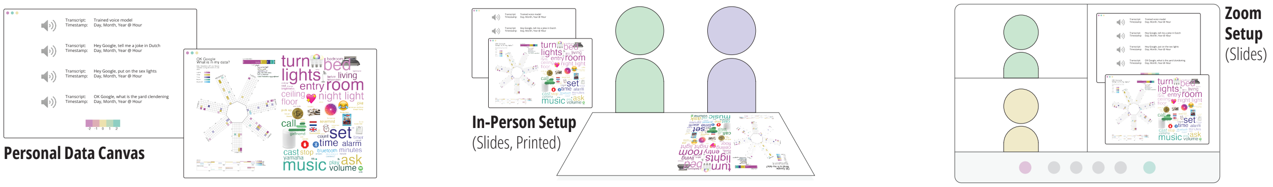
Published in
CHI 2023 - Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems

What is Sensitive About (Sensitive) Data? Characterizing Sensitivity and Intimacy with Google Assistant Users

Alejandra Gómez Ortega
Delft University of Technology
Delft, The Netherlands
a.gomezortega@tudelft.nl

Jacky Bourgeois
Delft University of Technology
Delft, The Netherlands
j.bourgeois@tudelft.nl

Gerd Kortuem
Delft University of Technology
Delft, The Netherlands
g.w.kortuem@tudelft.nl



Transcript: Trained voice model
Timestamp: Day, Month, Year @ Hour

Transcript: Hey Google, tell me a joke in Dutch
Timestamp: Day, Month, Year @ Hour

Transcript: Hey Google, put on the sex lights
Timestamp: Day, Month, Year @ Hour

Transcript: OK Google, what is the yard clendening
Timestamp: Day, Month, Year @ Hour

Figure 1: Example of the personal data canvas and setup of the in-person and online data-centric interviews. Personal data canvas shown with permission of the donor.

ABSTRACT

Digital technologies have increasingly integrated into people's lives, continuously capturing their behavior through potentially sensitive data. In the context of voice assistants, there is a misalignment between experts, regulators, and users on whether and what data is 'sensitive', partly due to how data is presented to users; as single interactions. We investigate users' perspectives on the sensitivity and intimacy of their Google Assistant speech records, introduced comprehensively as single interactions, patterns, and inferences. We collect speech records through data donation and explore them in collaboration with 17 users during interviews based on predefined data-sharing scenarios. Our results indicate a tipping point in perceived sensitivity and intimacy as participants delve deeper into their data and the information derived from it. We propose a conceptualization of sensitivity and intimacy that accounts for the fuzzy nature of data and must disentangle from it. We discuss the implications of our findings and provide recommendations.

CCS CONCEPTS

• **Security and privacy** → **Human and societal aspects of security and privacy**; • **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

Voice Assistants; Personal Data; Sensitive Data; Intimate Data;

ACM Reference Format:
Alejandra Gómez Ortega, Jacky Bourgeois, and Gerd Kortuem. 2023. What is Sensitive About (Sensitive) Data? Characterizing Sensitivity and Intimacy with Google Assistant Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3544548.3581164

CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581164

1 INTRODUCTION

People interact daily with products and services that collect personal data. Personal data is defined in the European General Data Protection Regulation (GDPR) as data through which a person can be directly or indirectly identified [23, Art. 4]. One example of products that collect personal data is voice assistants, embedded in ubiquitous devices that people interact with daily, including smartphones, smart speakers, smartwatches, and cars. Users of voice assistants interact with them through their voice, which is considered a convenient and natural way to communicate, more intuitive than clicking or typing [55]. In doing so, users integrate these devices into their routines and physical spaces, including pockets, bedrooms, living rooms, and kitchens [11, 47, 54]. Voice assistants collect personal data that is (1) volunteered, explicitly created and shared by a person (e.g., name and date of birth when filling out a registration form); and (2) observed, implicitly collected and captured by recording the actions and behavior of a person (e.g., timestamped speech records and textual transcriptions, generated and stored from each interaction) [14, 64]. Thus, voice assistants collect personal data containing various information about a person, some potentially sensitive.

The GDPR defines sensitive data as a special category of personal data that includes racial or ethnic origin, political opinions, religious or philosophical beliefs; trade-union membership; genetic data, biometric data processed solely to identify a human being; health-related data; and data concerning a person's sex life or sexual orientation [23, Art. 9]. Outside of the GDPR, the term sensitive data is used more broadly by Human-Computer Interaction (HCI) scholars, referring to information that is stigmatized (e.g., mental illness [7], HIV status [62]), should not be disclosed [34], could be easily compromised if disclosed [61], or whose disclosure could expose people and lead to inferences about their behavior and experiences [33, 52]. In addition, HCI scholars have introduced the term intimate data referring to personal information about intimate practices (e.g., cooking, sleeping, showering) taking place in intimate spaces (e.g., home) [27, 30] and bodily experiences (e.g., menstruating, urinating) [5, 28].

Speech records are collected in intimate spaces, namely people's phones and homes [30, 56]. They include data from all kinds of interactions, most of which are simple and mundane (e.g., "OK Google, set an alarm"[1]) [11, 47, 54]. Do they contain 'sensitive' or 'intimate' data? Let us say a person interacts with a Google Assistant while showering, "OK Google, play music on Spotify", could this be sensitive or intimate information? If so, why? Is it her voice? The voice is biometric information and is considered sensitive under the GDPR if used to identify a person. Is it the content? The context? She is in the shower, an intimate space. Is it the aggregation of multiple interactions that could lead to inferences about her showering routine?

Privacy experts and regulators argue that speech records correspond to sensitive information (especially the audio recordings) and emphasize that sensitive information about people's behavior can be inferred from them [43, 47]. Yet, voice assistant users seem to express a contrasting opinion. Previous research has shown that voice assistant users consider that individual speech records do not correspond to sensitive information [31, 36]. These studies focused on single interactions people had with their voice assistants, asking them about the acceptability of randomly selected interactions [36], and prompting them to reflect on interactions logged in a diary [31]. Speech records illustrate ambiguities around the interpretations of sensitive data. These are partly due to the focus on data as single interactions, disregarding data aggregation and potential inferences, of which voice assistant users have limited understanding [18, 31, 47]. Thus, there is a need to investigate how people articulate the notions of 'sensitivity' and 'intimacy' of speech records when introduced comprehensively, considering patterns and inferences in addition to single interactions.

In this paper, we investigate people's perspectives of 'sensitive' and 'intimate' data in the context of Google Assistant speech records. We aim to determine whether people are concerned about their speech records, the information they capture, or its potential disclosure; and what about speech records makes them 'sensitive' and 'intimate' data. Hence, we collaborate with Google Assistant users to address the following research question: **How do voice assistant users perceive 'sensitive' and 'intimate' data when faced with a comprehensive view of their speech records?** Specifically, we aim to understand:

(RQ1) What are the characteristics of speech records in terms of 'sensitivity' (as defined in the GDPR) and 'intimacy' (as understood in HCI)?

(RQ2) How do users articulate sharing their speech records?

(RQ3) How do users articulate speech records as 'sensitive' or 'intimate' data?

Our research is grounded in the speech records of 22 Google Assistant users we obtained through data donation [24, 58]. To develop a comprehensive understanding of speech records, we analyzed the dimensions and characteristics of the received data in relation to previous literature. We mapped them in terms of existing notions of sensitivity (as defined in the GDPR) and intimacy (as understood in HCI research). The results of this analysis fed the development of a scenario-based interview protocol around data-sharing and sensitivity. Of 22 users who donated their data (i.e., donors), 17 volunteered to participate in these scenario-based, data-informed interviews. We designed our study to allow donors to choose how they want to engage, resulting in 5 of them donating their speech records without participating in the interview. Throughout the interviews we focused on single interactions, patterns, and inferences. This process led most donors to experience a tipping point where their perceptions of the sensitivity and intimacy of their speech records changed. Our findings suggest that sensitivity is associated with the disclosure of information that is (1) intrusive, (2) specific, and (3) (un)available, while intimacy is a subset of sensitivity and unfolds through the recording of intimate thoughts and activities. We propose a conceptualization of 'sensitive' and 'intimate' data which accounts for people's perspectives. We discuss implications for data holders, policy makers, and researchers involved with 'sensitive' and 'intimate' data. Additionally, we present recommendations for researchers aiming to support research participants in disentangling the sensitivity and intimacy of their speech records.

2 BACKGROUND

2.1 Speech Records

Millions of people around the world use voice assistants. In 2022 Google Assistant and Apple's Siri each had over 500 million users worldwide, while Amazon's Alexa had over 100 million users worldwide[2]. Users of voice assistants integrate these devices into their daily and social life at home and outside the home [11, 47, 54]. Interactions start with a wake word, "OK Google", "Hey Siri", or "Alexa", which the always-on [36] voice assistant recognizes, processes, and stores. Hence, every interaction generates a speech record, which contains a timestamp, indicating the date and time, a transcript and an audio recording[3]. Thus, speech records allow for a detailed picture of voice assistant users and their routine activities [11, 47, 54]. This picture can include, for example, how someone lives in or near a city ("OK Google, what is the weather like in [city]?"), wakes up early in the morning during the week ("OK Google, set an alarm for 6:00"), and plays a podcast on Spotify first thing in the morning ("OK Google, open Spotify").

Research on the field of user intent mining (e.g., [17, 50, 55]) has been used to categorize these interactions. Broder [17] proposes a taxonomy of web search with three key needs that are also applicable to interactions with voice assistants: (1) informational, where the purpose is to obtain information, (2) transactional, where the purpose is to perform an activity (mediated by the voice assistant), and (3) navigational, where the purpose is to invoke a third party application. Similarly, Qu and colleagues [50] propose a taxonomy that includes follow-up questions and greetings/gratitude. While

[1] The examples of interactions we present in this paper come from the 8735 speech records we received.

[2] Voice assistant users worldwide, from smart speakers global market report (accessed in September 2022)

[3] As of 2020 the Google Assistant only stores the audio recording if the user has opted-in.

Shani and colleagues [55] argue for considering playful interactions as well (e.g., "OK Google, tell me a joke"), and propose a taxonomy of playful interactions including relief (e.g., "OK Google, order poop"), incongruity (e.g.,"OK Google, give me a high five"), and superiority (e.g., "OK Google, you suck").

Due to the nature of interactions with voice assistants and where they take place, speech records could contain and lead to potentially sensitive and intimate information about users. Yet, previous research (e.g., [31, 36, 37, 40, 43]) illustrates nuances around perceived sensitivity and intimacy. For instance, a longitudinal study of ubiquitous surveillance in the home concluded that people consider audio recordings among the two most sensitive and disturbing data, the other being video recordings [43]. Similarly, researchers have documented privacy concerns around voice assistants in the home, including that they listen 24/7, can record private conversations, and collect personal information [19, 31, 37, 40]. In contrast, Malkin and colleagues [36] found that *'on the whole, data currently stored with voice assistants is not considered sensitive'* by users. Their research is limited to individual interactions (i.e., a single speech record) and does not consider what could be inferred from multiple interactions. Lau and colleagues [31] conducted a diary study with voice assistant users and obtained similar results, *'users did not consider their speech records sensitive and did not make use of privacy controls'*.

Concluding. There is a disagreement around perceptions of sensitivity that aligns with the limited awareness and understanding that users have of the content of speech records over time and the potential inferences that could be made from them, which is well documented in the literature [18, 36, 47]. We aim to investigate users' perceptions of 'sensitivity' when providing a comprehensive representation of their speech records, considering patterns and inferences in addition to single interactions. The latter has been the only focus of previous research (e.g., [31, 36]).

2.2 Personal, Sensitive and Intimate Data

The GDPR introduces specific categories of personal data that are considered sensitive data (Introduced in Section 1). Outside of data protection regulations and privacy laws, the term sensitive data has been used in a broader sense, referring, for example, to sensitive information that can be extracted from a seemingly innocuous dataset [52, 65]; or to the creepiness of learning about the data collection and distribution practices of digital products and services in different contexts [56]. HCI scholars have used the term sensitive data in various ways. For Liu and colleagues [34], who investigate sensitive data transfer, data is sensitive as it contains private information that should not be revealed. In the context of selective disclosure of data, Rudnicka and colleagues [52] understand sensitive data as information that could potentially be used to learn about a person and her routines, including her location, and health (included in the special categories of the GDPR). HCI scholars have also referred to specific types of data that are sensitive, such as financial information [61], mental and physical health [1, 7] and sex [32, 62] (the last two included in the special categories of the GDPR). Moreover, HCI scholars have introduced the term intimate data to describe data generated from intimate activities and contexts [5, 30]. These include everyday activities that take place in the home, such as cooking, sleeping, and showering [30]; as well as bodily experiences like menstruating and urinating [5, 10, 28, 39]. Intimacy in HCI also relates to physical togetherness [27, 45].

Speech records are personal data as they relate to a person. They contain potentially sensitive data (as defined in the GDPR), such as the voice. In addition, they can be reused to learn potentially sensitive information about a person, which could be misused (sensitivity as understood in HCI). From the audio recordings alone, it is possible to estimate the age of the speaker [59], recognize her emotions [19, 42], identify activities such as laughing, crying, and eating [19, 51], diagnose a broad range of psychiatric disorders, including depression and schizophrenia [19, 35], and determine the size and shape of the room where the device is located [19]. In addition, speech records generate in a shared physical space, where there is a physical closeness between a user and their voice assistant [27, 45]. Hence, they could contain intimate data from the intimate activities occurring within that space.

Concluding. When considering existing definitions of 'sensitive' and 'intimate' data, the personal information collected, stored, and inferred via speech records could be 'sensitive' and 'intimate'. Yet, as described in Section 2.1, users' perceptions differ. In this paper, we use the GDPR's definition of 'sensitive data' and the HCI conceptualization of 'intimate data' as a starting point to examine speech records and contrast them with users' perceptions.

2.3 Privacy and Contextual Integrity

HCI scholars have approached privacy through different lenses. Crabtree and colleagues [20] provide an overview of the various ways in which privacy is understood throughout the HCI literature, including privacy as control, privacy as boundary management, and privacy as contextual integrity, among others. Privacy as control relates to the ability to control the flow of personal data through activities such as limiting information disclosure [62] and filtering what gets disclosed [53]. Privacy as boundary management is informed by the work of Irwin Altman [6]; and relates to the selective disclosure of personal information as people move between privacy and publicity according to the context and intention [44]. Based on the boundary metaphore, Sandra Petronio proposes the Communication Privacy Management (CPM) theory; in which the disclosure of private information is based on privacy rules that are negotiated around personal and collective boundaries [46]. Privacy as Contextual Integrity (CI) is a theory proposed by Helen Nissenbaum where privacy is understood in terms of the appropriateness of information flows according to social or cultural norms and grounded in specific contexts [41].

CI is considered an appropriate framework to understand people's privacy norms and has been operationalized through large-scale surveys in different contexts [2, 8, 38, 57]. As proposed by Nissenbaum [41], information flows are described according to five parameters: (1) subject of the information, (2) sender of the information, (3) attribute, describing the type of information, (4) recipient of the information, and (5) transmission principle, stating the condition under which the information flow is permitted. For instance, a Google Assistant user (subject) might be comfortable with Google employees (recipient) reviewing the audio recordings (attribute) from her Google Assistant (sender) if she has opted-in for the collection and revision of voice and audio (transmission principle); but not with the police doing so (a different recipient and privacy violation). CI surveys inquire about the acceptability of information flows, illustrated through scenario-based vignettes with varying parameters (e.g., a different recipient or transmission principle). For instance, Apthorpe and colleagues explored the acceptability of information flows in smart home IoT devices, including fitness trackers, thermostats, and personal assistants, among others [8]. While Abdi and colleagues explored the acceptability of information flows in smart home personal assistants and considered several types of data, including voice recordings [2].

Concluding. In this paper, we draw inspiration from how CI has been operationalized in previous research [2, 8, 38, 57] to inquire about the acceptability of sharing speech records and its perceived sensitivity across predefined scenarios. We use these scenarios, grounded in personal data, as prompts to elicit reflection on 'sensitivity' and 'intimacy'.

3 METHODOLOGY

In this study, our goal was to investigate people's perspectives of 'sensitive' and 'intimate' data in the context of Google Assistant speech records. In particular, we sought to understand:

(RQ1) What are the characteristics of speech records in terms of 'sensitivity' (as defined in the GDPR) and 'intimacy' (as understood in HCI)?
(RQ2) How do users articulate sharing their speech records?
(RQ3) How do users articulate speech records as 'sensitive' or 'intimate' data?

Given the nature of the questions to answer, we aimed to gather speech records from voice assistant users interacting with their devices in the naturalistic setting of their routine. In addition, we aimed to actively support and involve users in obtaining a copy of and comprehensively exploring their data.

In this section, we describe our research activities. First, we collected speech records generated in-the-wild through data donation (Section 3.1). Then, we conducted an initial analysis of the received speech records (Section 3.2). Finally, we used the results of the analysis to define a data-centric interview protocol, which we conducted with 17 donors (Section 3.3). Our institution's Human Research Ethics Committee and Privacy Team reviewed and approved these activities.

3.1 Collection of Speech Records through Data Donation

We conducted a data donation campaign [24, 58] to collect speech records generated in-the-wild. Data donation enabled us to access data already generated by users during the routine interactions with their voice assistant outside the context of our study, as opposed to prompting participants to interact with their voice assistants in a specific way or to collect data. We decided to focus on Google Assistant users, as Google has a relatively simple and quick process to obtain a takeout of the data, an extensive pool of users, and well-structured metadata. Thus, between April and June 2022, we reached out to Google Assistant users (e.g., Assistant App, Google Home, Google Nest) worldwide, and we invited them to donate their speech records and participate in an optional interview. We used snowball sampling by periodically posting our 'call to donate' across different channels for three months. These included our personal social media (e.g., Twitter and LinkedIn), online communities (e.g., subreddits r/GoogleAssitant and r/GoogleHome, Google Home users on Nextdoor), and local cafes and universities. Additionally, we reached out to people and institutions (e.g., a privacy foundation and an internet podcast) who shared our 'call to donate' with their communities through social media, newsletters, mailing lists, and events in which we took part. In total, our 'call to donate' had an estimated of 35.000 views. We asked interested users to download a copy of their speech records from Google, upload it to our data donation platform[4], and decide whether to participate in the interview. Due to Google's 2020 policy change requiring users to opt into voice data collection, four donors had to opt in and collect data for a couple of months before donating it. These donors knew their interactions with their Google Assistant would be used for our research.

3.1.1 Participants: Data Donors.

22 users of Google Assistant (referred to in the paper as donors D1-D22) volunteered to participate in our research by donating their speech records (N = 22, 1 identified as non-binary, 7 as female, and 14 as male). They ranged in age from 21 to 58 years (mean = 30.8, median = 38). Out of these, 17 ($n_{interview}$ = 17, 5 identified as female and 12 as male) agreed to participate in the interview. Donors were located in the European Union (EU) and South America. Obtaining a copy of the speech records, enabled by the GDPR, was also possible for donors outside the EU[5].

3.2 Initial Analysis: Familiarization and Classification

Throughout the data donation campaign, the first two authors independently analyzed the donated speech records to gain insights into common queries and patterns within individual datasets[6]. The purpose of this activity was twofold: (1) identify the characteristics of speech records and (2) recognize the relevant attributes to structure the data-centric interviews. In total, we received 8375 speech records (i.e., individual interactions with a Google Assistant), although the number of speech records obtained per donation varied widely. It depended on how long and how often donors interacted with their Google Assistant. The largest dataset contained 5766 speech records, and the smallest had 24. Although one donor contributed a significantly larger dataset, this did not influence how we used and analyzed speech records in this study, i.e., as prompt and support for qualitative exploration. We do not derived quantitative insights from the donated speech records; other than a descriptive overview. In addition, we found the same types of interactions and information in each of the datasets received. Therefore, we present the same type of prompts during the interviews, based on

[4] Data donation platform: working prototype in datadonation.ide.tudelft.nl and data storage and sharing source code in github.com/datacentricdesign/bucket.
[5] The GDPR applies to the population of the European Union. Yet, in practice, the right to data portability is available worldwide, since international companies rarely limit it by geography [14]. The Google Takeout dashboard (takeout.google.com) is available to users worldwide.
[6] We provide the protocol of our analysis in the supplementary material.

Intended Interactions				Unintended Interactions	
Informational Request	**Navigational Request**	**Transactional Request**	**Playful Request**	**Background Conversation**	**Third Party Recording**
How long does it take me to go to [address]? Inferences (Sensitive HCI) Indirect Location and Habits	*Open YouTube* Inferences (Sensitive HCI) Preferences and Habits	*Turn off the office lights* Inferences (Sensitive HCI) Equipment and Habits	*Tell me a joke*	*Unintelligible conversation at home*	*Unintelligible movie fragment*
What does it mean if you have [health condition]? Health-Related (Sensitive GDPR)	*Play [song] by [artist] on Spotify* Inferences (Sensitive HCI) Interests and Habits	*Turn on sex mode in the living room* Sex-Related (Intimate HCI) (Sensitive GDPR)	*Tell me a joke about Donald Trump* Political Information (Sensitive GDPR)	*Intelligible conversation with a public official*	*Intelligible movie fragment* Inferences (Sensitive HCI) Interests

Figure 2: Example of initial analysis. The speech records (*italic*) are categorized and annotated (red) with respect to sensitivity and intimacy.

their personal datasets, to donors. The first two authors systematically listened through each speech record and classified them as intended (i.e., the speaker used the activation command) or unintended (i.e., the speaker did not use the activation command). We then mapped the intended interactions into one of four categories. These came from Brode's web search taxonomy [17], integrating playfulness, common among voice assistant users [11, 31, 55]. The categories were: (1) informational request (e.g., "OK Google, what's the weather like?"), (2) transactional request (e.g., "OK Google, turn on the lights"), (3) navigational request (e.g., "OK Google, open Spotify"), and (4) playful request (e.g., "OK Google, tell me a joke"). Additionally, we mapped unintended interactions into one of two categories, according to the content of the audio recordings: (1) (background) conversation and (2) third-party audio (e.g., TV or radio). Finally, we determined if each speech record contained sensitive (as defined in the GDPR or understood in HCI) or intimate (as understood in HCI) information and annotated it. When sensitive or intimate information derives from inferences (e.g., "OK Google, How long does it take me to go to [address]?"), we annotated potential inferences (e.g., goes to [address] every Monday) considering similar interactions throughout the dataset and the state of the art on the different information derived from interacting with personal voice assistants (e.g., [19]). Figure 2 illustrates this process.

3.3 Data-Centric Interviews

Of 22 donors, 17 volunteered to participate in the interviews. Hence, we conducted semi-structured interviews with 17 donors, prompted by their donated speech records. The interviews aimed to facilitate donors' reflection and exploration of their speech records and to capture thick and nuanced insights about sharing, sensitivity, and intimacy. The first author conducted the interviews in English between June and July 2022. Interviews lasted between 35 and 55 minutes; 5 took place in person and 12 via Zoom. The first author conducted one interview with the two members of a household who share a device ($D9_{a,b}$), the remaining 16 interviews were one-on-one as most donors were single-users of their Google Assistant. During the interview, we aimed to comprehensively explore speech records; hence we delved into individual speech records (perceived as not sensitive by users [31, 36]) as well as patterns and inferences derived from multiple speech records (not considered in previous literature [31, 36, 47]). The first author designed a personal data canvas for each interviewee (Fig. 3) containing individual interactions as audio clips (Fig. 3a) and a visualization of the data (Fig. 3b) to support the interviews.

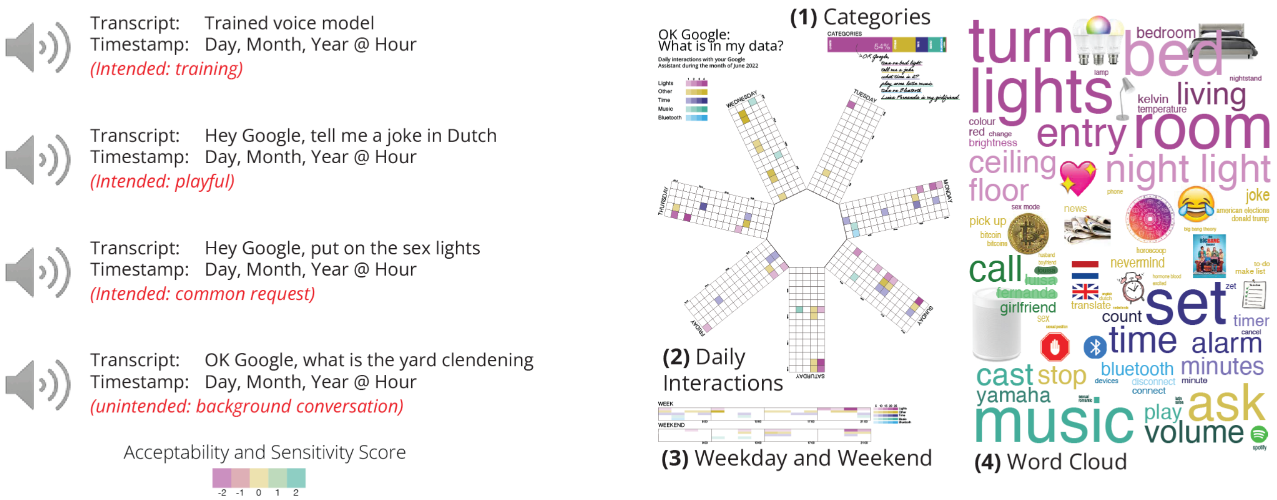
3.3.1 Interview Protocol. Inspired by the Contextual Integrity (CI) scenario-based inquiry, the interviews revolved around six attributes introduced with examples from each interviewee's dataset and presented through the personal data canvas[7]. Specifically, we focused on introducing the different types of single interactions present across the received datasets. These encompass a wide range of use cases and contexts: (1) a neutral and de-contextualized training interaction; (2) a simple yet telling playful interaction; (3) a common request interaction (informational, navigational, or transactional), illustrating common usage patterns and contexts of use; and (4) an unintended interaction that signals an unexpected device operation. In addition, we focused on introducing the main types of information that can derive from the aggregation of multiple interactions: (5) patterns and (6) inferences. These attributes are not mutually exclusive categories, but rather illustrative examples of various information grounded in the received datasets. We selected these attributes to provide a comprehensive overview of the speech records and the underlying information they contain, derived from the initial analysis (Section 3.2). For each attribute, donors answered two questions:

(Q1) How acceptable is it for you to share **<attribute>** with **<recipient>**?

(Q2) How sensitive do you consider this **<attribute>** to be?

We introduced Q1 since sensitivity is often associated with information disclosure and governing information flows (e.g., [34, 52, 61, 62]). Hence, prompting participants to consider what information to disclose and with whom invites them to reflect on the characteristics of that information. For Q1, the recipients included: **partner(s), family, friends, colleagues, and researchers**. We selected this list to investigate how the acceptability to share varies within one's extended personal network and its implications regarding sensitivity and intimacy. Generally, CI scenario-based inquiries explore and vary five parameters: sender, attribute, recipient, and transmission principle. We vary only the attributes and recipients since in our wording of the question we emphasize people's agency over the transaction. Additionally, we wanted to reduce the complexity

[7] We provide the interview protocol in the supplementary material.



Transcript: Trained voice model
Timestamp: Day, Month, Year @ Hour
(Intended: training)

Transcript: Hey Google, tell me a joke in Dutch
Timestamp: Day, Month, Year @ Hour
(Intended: playful)

Transcript: Hey Google, put on the sex lights
Timestamp: Day, Month, Year @ Hour
(Intended: common request)

Transcript: OK Google, what is the yard clendening
Timestamp: Day, Month, Year @ Hour
(unintended: background conversation)

(a) Single interactions. **(b) Multiple interactions.**

Figure 3: Example of a personal data canvas. Shown with permission of the donor.

for donors by limiting the parameters and length of the question. Therefore the subject (i.e., user), the sender (i.e., Google Assistant prompted by its user), and the transmission principle (i.e., the user is informed and notified) remain constant. We introduced these at the beginning of the interview and brought them up in case further clarification was needed. We invited the donors to answer Q1 and Q2 using a 5-point Likert-Scale[8]. We computed the average acceptability and sensitivity scores based on these [8, 9]. In addition, we invited donors to articulate the rationale behind their choice and elaborate on their responses. We did not want to impose the term 'intimacy'. Hence, we deliberately refrained from introducing it and we only discussed it if and after it was introduced by donors. In addition, we deliberately left open the definitions of 'sensitivity' and 'intimacy' as we wanted donors to express themselves on their terms and bring their interpretations.

3.3.2 Personal Data Canvas. The personal data canvas is a representation of the donor's data (Fig. 3)[9]. We opted for visualizing the data as it has been a successful approach to conduct interviews prompted by data and support interpretation and reflection [12, 13, 29, 36]. We designed it with the objective to comprehensively introduce the speech records and bring focus to the specific attributes. It consisted of two interactive views presented on a screen where donors could click and reproduce audio clips and zoom in and out of the different visualizations. If the interviews were in person, the visualization (Fig. 3b) was also printed on A3 paper, inviting donors to explore it from multiple angles and annotate it. In the first view (Fig. 3a), we focused on single interactions (e.g., "OK Google, turn on the lights"). Specifically, we presented the following interactions: (1) training (e.g., "OK Google", generated when a user is first configuring their device and 'training' it to recognize her voice), (2) playful, (3) common request and (4) unintended. For these, we presented a clickable button with the audio recording (which is reproduced), and its transcript and timestamp in a human-readable format. In the second view (Fig. 3b), we focused on multiple interactions, where the focus lies on the combination of timestamps and transcripts leading to potentially sensitive information, namely (5) patterns in time or sequence and (6) inferences. For these, we visualized the data and invited donors to explore the visualization, reconstruct the context of the data, and reflect on their patterns and potential inferences.

In the visualization (Fig. 3b), we focused on conveying the information from the timestamps and transcripts of multiple interactions. Specifically, we identified common interactions for each dataset and grouped them into categories (e.g., weather, music, time). We visualized the distribution of these categories throughout the dataset with a bar graph (Fig. 3b$_{(1)}$), and we represented each category with a different color throughout the visualization. Additionally, we presented the number of daily interactions for each category per hour of the day and day of the week with a heat map (Fig. 3b$_{(2)}$) where we focused on the 16 hours of the day with more interactions, the start and end times vary by donor. Similarly, we used a heat map to present the number of interactions of each category per hour of the day during the weekdays (Monday through Friday) and weekends (Saturday and Sunday) (Fig. 3b$_{(3)}$). Finally, we presented a word cloud (Fig. 3b$_{(4)}$) with the most frequent words grouped and color-coded by category, additional images were visually representing some of the terms. We added the images to make the interactions more prominent and easier to explore.

3.3.3 Reflexive Thematic Analysis. The interviews were audio recorded and transcribed. The first author made an initial transcript using

[8] Likert-Scale, Q1 from completely unacceptable (-2) to completely acceptable (+2), and Q2 from completely sensitive (-2) to completely not sensitive (+2)
[9] We provide a full-size example in the supplementary material.

MS Office 365, then manually reviewed and edited it. The first two authors analyzed the transcripts using reflexive thematic analysis [15, 16], within a constructionist framework. Both authors independently read through the transcripts to familiarize themselves with the data and coded the entire dataset using ATLAS.ti. Through this process, we aimed to capture all the aspects of the data relevant to understanding how data subjects articulate sharing, sensitivity, and intimacy of their speech records. Both authors independently reviewed the codes and subsequently discussed and grouped them into tentative themes. The first author iteratively reviewed and refined the themes.

4 RESULTS

In this section, we introduce the characteristics and dimensions of the speech records determined from the initial analysis (Section 4.1); and we present the findings of the reflexive thematic analysis of the data-centric interviews regarding sharing, sensitivity, and intimacy (Sections 4.2 and 4.3). The examples of interactions we present throughout this section come from the 8735 speech records we received.

4.1 Characteristics and Dimensions of Speech Records

4.1.1 Characteristics. Speech records contain brief interactions between a person and her Google Assistant, generally in the form of a request (e.g., "OK Google, set timer") or a reply to a request (e.g., "thank you"). On average, the speech records we received had 4.40 words per interaction – excluding the wake word; and lasted between 2 and 6 seconds. Donors interacted with their Google Assistant on average 3.91 times a day, with most interactions taking place in the early morning, late afternoon, and evening (Fig. $4_{(5)}$). The most frequent interactions were about turning on and off the lights, alarms, and reproducing music on a third-party app or device. These findings are consistent with previous research on the long-term use of voice assistants [11]. We classified a small percentage (1.05%) of the interactions we received as unintended; most of these were recordings of background conversations. The majority of the interactions (98.95%) corresponded to intended interactions, as they were initiated from the wake word or were part of a series of intended interactions (Fig. $4_{(1)}$). From these, more than half (62.54%) corresponded to transactional requests, followed by navigational (16.86%), informational (15.05%), and playful requests (1.54%) (Fig. $4_{(2)}$). Furthermore, in all the datasets, more than one person (i.e., speaker) was present in the audio recordings.

Each type of request contains different layers of information about the person interacting with the Google Assistant. Transactional requests contain information about a person's set-up and the Google Assistant capabilities they use (e.g., "OK Google, set bed light to rainbow", "OK Google, change the blood pressure pill reminder time to 8 am"). Transactional requests potentially contain information about people's routines and habits from which information regarding a person's health, political interests, and sex life could be derived, considered sensitive under the GDPR. Navigational requests contain information about a person's digital routine and habits, including the digital content they consume and the third-party apps they interact with (e.g., "OK Google, play [song]", "OK Google, start instrument tuner"). These do not contain any information considered sensitive under the GDPR. Informational requests contain information about a person's interests and concerns; including questions about current affairs (e.g., "OK Google, who is the prime minister in [country]?"), relevant events and activities (e.g., "OK Google, when does the new electricity rate apply?", "What is the wheel to pray the rosary called?"), people's bodies (e.g., "OK Google, how do I know if I injured my rotator cuff?"), and even dreams (e.g., "OK Google, what does it mean to dream that someone dies?"). Several categories of sensitive information can be derived from informational requests (e.g., political opinions, religious beliefs, health- and sex-related). In addition, the low threshold of interacting with the Google Assistant means speech records explicitly include sensitive and intimate questions (e.g., "OK Google, why did I get so dizzy after [medical procedure]?", "OK Google, what is the sexual cowgirl position?").

The distribution of request types varies between people who use Google Assistant on a smart speaker (13 donors) and those who use it on a smartphone (9 donors) (Fig. $4_{(5)}$). In the first scenario, transactional requests are more frequent, as smart assistants are often integrated and connected with other smart appliances, while in the second, informational requests are more frequent (Fig. $4_{(4)}$).

4.1.2 Dimensions. We introduce the three dimensions of speech records: *timestamp*, *transcript*, and *audio recording* by providing a short description and illustrating the information that can be derived from each.

- *Timestamp*, the date and time of an interaction (precision of milliseconds), describes when actions and interactions take place. For example, asking Google to "set an alarm for 8 am" late at night can indicate when a person goes to sleep and when she intends to wake up. When multiple interactions are combined, the timestamp can illustrate patterns in time (e.g., snoozes the 8 am alarm every weekday) and sequence (e.g., snoozes the 8 am alarm, plays a news podcast, asks about the weather). It can highlight specific aspects of a person's routine. Moreover, the timestamp enables data to be interpreted and abstracted according to different instants, such as the time of day (e.g., the middle of the night) and the month of the year (e.g., September). These can be associated with external factors (e.g., daylight, weather, public holidays) or situated within larger contexts (e.g., pandemic).
- *Transcript*, the content of the interaction, as interpreted by Google Assistant, describes what an interaction is about. Transcripts can reflect a person's worries (e.g., "OK Google, what should I do to protect myself from Corona?") and interests (e.g., "OK Google, how did [soccer team] score this season?"). Transcripts can also indicate a person's location (e.g., "OK Google, what's the weather like in [city]?", "OK Google, movies in [cinema], [city] today"). Moreover, transcripts illustrate people's relationships (e.g., "OK Google, [name] is my girlfriend", "OK Google, call my mom") and how people relate to their Google Assistant (e.g., "OK Google, set reminder for tomorrow morning, *darling*", "OK Google, you are so *stupid*"). The inferences derived from the transcripts are limited by and specific to the way people interact with their Google Assistant.

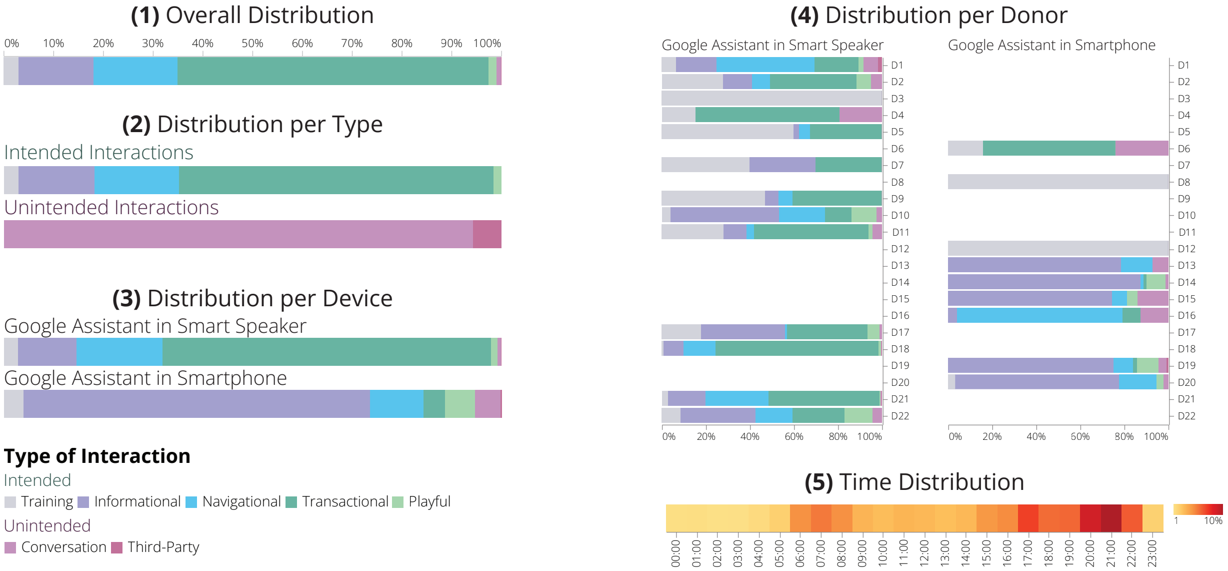


Figure 4: Distribution of (1) types of interaction across the received datasets (2) types of intended and unintended interactions (3) types of interaction when using the Google Assistant on a smart speaker and smartphone (4) types of interaction in the datasets received from each donor (5) interactions by the time of day over the entire data set.

- *Audio Recording*, The sounds and speech of an interaction, as recorded by the Google Assistant in an audio file, describes who (i.e., speaker) and how (e.g., quiet room and sleepy voice, or loud background music and loud voice) interacts with the Google Assistant. Audio recordings can help differentiate intended and unintended interactions. The voice is a distinctive element in audio recordings, that identifies and distinguishes the speaker(s). The voice is rich and nuanced and communicates more than just words. Hence, interactions gain an extra layer when considering the nuances of the voice.

4.2 Sharing Speech Records

4.2.1 Average Acceptability Score. We calculate the acceptability score (Fig. $5_{(1)}$) by averaging the responses to Q1 (about sharing with different recipients) for each attribute and each recipient [8, 9]. Generally, donors considered it acceptable to share all attributes with most recipients, partner(s), family, friends, and researchers ($acceptability\,score \geq 1$). The acceptability score is lower, tending toward neutral ($0 < acceptability\,score < 1$), for colleagues, especially for single unintended interactions and the patterns and inferences derived from multiple interactions.

4.2.2 Boundaries. A recurring theme in our analysis is the boundary between people's private and public lives (and spaces), which shape what is acceptable to share with whom. It aligns with the conceptualizations of privacy as boundary management proposed by Altman [6] and Petronio [46] and it replicates some of their findings. We briefly introduce it as part of our results as it is highly relevant to the concepts of sensitivity and intimacy. D17 describes this boundary when referring to what can be inferred from her speech records:

> "Sometimes I ask pretty weird things to Google. So yeah, it shows a little bit more the weird parts of me, that I don't want to show everyone, other that, for example, my partner." (D17)

The acceptability to share personal data with a recipient decreases the further the recipient is from the context and the space in which data are generated. Sharing with recipients who are inside the private space is considered acceptable since they already inhabit that space and are familiar with what happens within, *"it is information that you usually talk* [about] *with your close people, that they already know. Your routine, your activities, if you are going somewhere, it is something that you share all the time"* (D16). In some cases, being inside this space means they are even part of the data, *"I mean he's there* [in the audio recording]*"* (D19).

Sharing with recipients who are in *the boundary* of the private space is considered less acceptable, *"friends, and colleagues, they don't have to know you are going* [somewhere]*"* (D10), *"they don't need to know what I do, what is my routine"* (D16). Because these recipients do not belong to the private space, they are unaware of its peculiarities *"they don't know the context"* (D18), and may misinterpret, misunderstand or judge what happens within, *"if I share that* [playful interaction] *with my colleagues, they can judge me, and I interact with them so the judgment is more real"* (D11). The concern of being judged comes from the lack of control over how data, incomplete and decontextualized, is received by others.

Sharing with recipients who are outside the private space is generally considered not acceptable. Here, data might lead to an opportunity to access and violate the private space, *"that's something that people can abuse"* (D21). D17, who has smart lights linked to her Google Assistant, describes the possibility of harm when discussing these interactions:

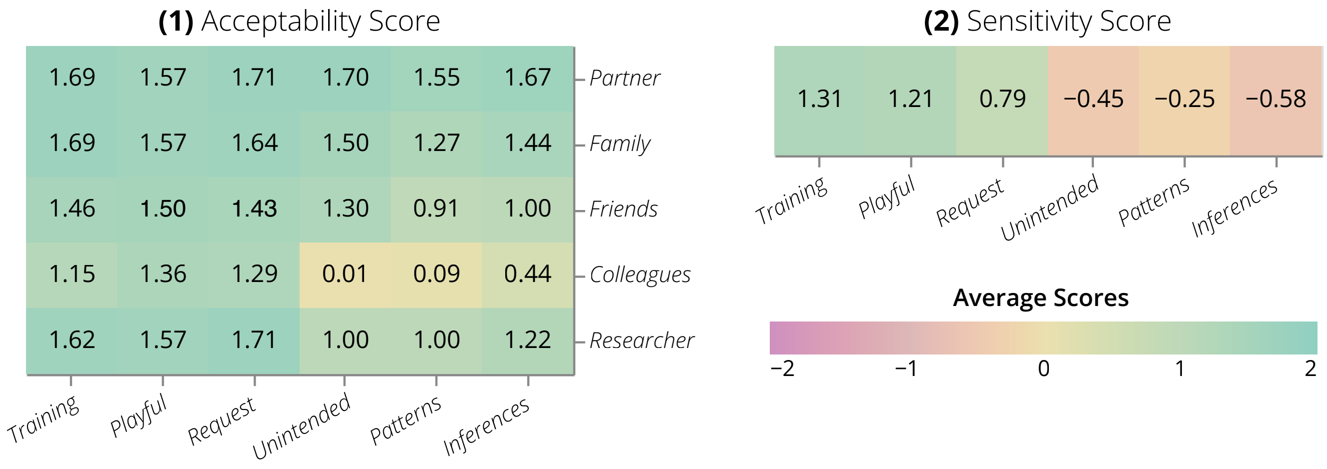


Figure 5: Average (1) acceptability and (2) sensitivity scores

> "[the light interactions] tell about my routines, when I'm home and when I am not home. So this might be used against me [...] If someone wants to threaten me or if they want to steal something in my house or something like that." (D17)

Researchers, who are outside the private space, are a notable exception. Donors highlighted the importance of contributing to science and advancing research, *"I believe in science, so they* [researchers] *may use my information"* (D17). Evidently, donors already shared their personal data with our research. For donors, researchers are responsible and carry values such as trust and discretion. As such, they are expected to make proper use of the data, *"handle this* [data] *with precaution"* (D1).

4.2.3 Characteristics of the Data. We developed three themes related to the characteristics of the data that are important to consider in the context of sharing, sensitivity, and intimacy. These are not specific to individual speech records or the information derived from them; hence, we use the encompassing term data.

- Data is *contextual.* It is generated and stored in a given context, which is lost when translated into a discrete event. For example, the interaction "OK Google, how does a guinea pig sound" occurs in a specific context, *"I use it to demonstrate how my guinea pigs react. And its purpose can be various. It can be for friends, it can be for yourself, it can be for your cats. It's extremely funny* [laughs]" (D10). Yet, it loses playfulness and social dynamics when recorded. Hence, data can *"always be misinterpreted"* (D18) if it is disassociated from its context. In addition, because of how the data is generated (e.g., through seamless voice interactions), it portrays a limited and specific image of a person that loses the nuance embedded in the context.
- Data is *relational.* It relates to the interaction between a person (or people) and a device, sharing a physical space (e.g., a room), as well as other people, not necessarily sharing a physical space (e.g., partners, family members, roommates, neighbors, visitors). In addition, it accounts for relationships with others, *"I'm putting information out there about our relationship"* (D21). In fact, more than one speaker was present in all received datasets, especially in multi-user environments, where more than one person shares space and device (e.g., partners, family members, roommates), but also in single-user environments where other people are occasionally around (e.g., neighbors, visitors).
- Data is *multiple.* A term introduced by Prainsack [49] referring to it being able to be and used in more places than one at the same time. It can be duplicated and shared, which became apparent during the data donation process when donors obtained and shared a 'copy' of their data. Multiplicity illustrates how data can be reused, and potentially misused, for more than one purpose and more than one entity. In addition, data when aggregated and combined becomes a combination of *"multiple fragments* [from which] *you can distill some meaning"* (D21).

4.3 Sensitivity of Speech Records

Donors interpreted the sensitivity of speech records in three ways. First, how personal or private individual speech records or the information derived from them are. Here, sensitive speech records contain personal information (from within the private space). Second, how disclosable individual speech records or the information derived from them are. Here, sensitive speech records are not to disclose; their disclosure would be considered a violation or potentially harmful. Third, to what extent can individual speech records or the information derived from them be expanded or used differently. Here, sensitive speech records can lead to inferences about personal information, from within the private space, and can be re-used outside the context in which they were generated. Starting from these interpretations, we describe the characteristics of sensitive speech records and map these into a sensitivity spectrum.

4.3.1 Characteristics. We developed three themes describing the characteristics of sensitive data (Fig. 6): *intrusiveness*, *specificity*, and *(un)availability.* Sensitive data is *intrusive* as it is generated within

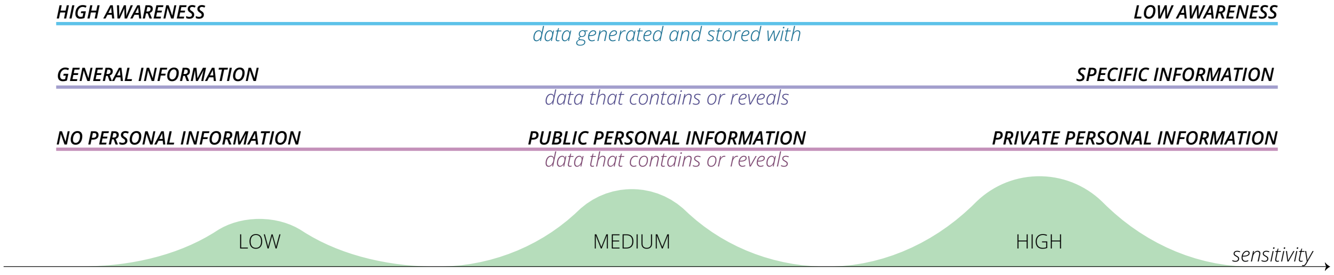

HIGH AWARENESS
LOW AWARENESS
data generated and stored with
GENERAL INFORMATION
SPECIFIC INFORMATION
data that contains or reveals
NO PERSONAL INFORMATION
PUBLIC PERSONAL INFORMATION
PRIVATE PERSONAL INFORMATION
data that contains or reveals
LOW
MEDIUM
HIGH
sensitivity

Figure 6: The spectrum of sensitive speech records.

people's private spaces where others are not necessarily welcome or invited. Intrusiveness unfolds through a shared physical space where, by capturing what happens within, others gain indirect access through data.

> "That I'm planning [activity] it's something that I believe should be in my private area. It's something that I don't feel comfortable seeing on that screen. Because it means someone has the opportunity to know what I'm planning, where I'm going, what I'm thinking, things that are worrying me, very, very, very personal things." (D14)

Intrusiveness is related to people's awareness of the device's data collection and storage practices and how they experience them. Hence, it manifests when the device captures more than expected.

> "It's like if you're naked you don't say «OK, Google» you first dress up, and then you say «OK Google», which is, I don't know if it is a lousy metaphor. But what I'm trying to say is that we feel aware that for a few seconds when we say «OK Google», it's not like, we're not exactly alone." ($D9_b$)

As expressed by $D9_b$ donors described a sense of awareness and control over their interactions with the Google Assistant that is lost when data is generated from unintended interactions or interactions that record *"a little bit more than my question"* (D1). Thus, the device intrudes further into a space by capturing data that reveals more information than people intended to disclose or were unaware of. Moreover, intrusiveness relates to the amount of data that is made available, as more data means more angles from which to access a person's private space and more possibilities of unawareness, *"anyone can have one digit of my pin. I'm happy to give everyone one digit of my pin. But not all four"* (D21).

Sensitive data is *specific* as it more narrowly records and reflects certain aspects or themes of a person's private space. D17 illustrates how sensitivity varies with more and less specific points when comparing two interactions, the first about her mom and the second about the office lights.

> "The [first] one attaches something to me as a person that is my mother. So, some information regarding something specific to an individual. This [the second] is really general; all kinds of people have office lights. When I ask about my mom, I know that the answer is going to have specific information in it, so that's why I consider it more sensitive." (D17)

Specificity relates to what is uniquely about a person, as opposed to generic information that can be attached to anyone or that relates to decontextualized activities.

> "It's about how unique is that data. Generalistic things that everyone does, that don't have any specific anchoring in space or time, for instance, the fact that I live in [city], it's on my profile. The fact that I live on that particular street, and like on Tuesday he does this, that is more sensitive, so it's about the specificity of the information." (D21)

In addition, specificity relates to data being spatiotemporal. This means data can reveal when and where certain interactions occur as they have a *"specific anchoring in space or time"* as described by D21. It gives an extra layer of specificity to the data, as it can be further situated by considering external factors such as the day of the week and the weather. Furthermore, specificity may derive from the aggregation of multiple data and sources of data, resulting in a distinct representation of certain aspects of a person's private space, *"they know, hey, that is* [name] *and combined with all the things you do on the internet. Well, it gets pretty sensitive"* (D10).

Sensitive data is *(un)available*, as it is generated within a person's private space where *"stuff is much more personal. So I want to guard it a little bit more"* (D11). For instance, the health-related questions that people ask Google before even discussing them with their family, friends, and doctors.

It is collected and stored, hence it is potentially available, yet it is generally not at someone's disposal. It reveals information that is not widespread or publicly available as opposed to information that people *"probably can find everywhere"* (D18). The increasing availability of personal data online and on social media means that widespread information from within the public space is not necessarily considered sensitive.

> "I have social media, so yeah, it's normal that these things are known by a lot of people that I don't know about" (D17).

The combination of multiple interactions over time means sensitive data, in the form of patterns and inferences, is sometimes (un)available even to the data subjects.

> "You know? It's weird. Like I think I'm not conscious of how much they know about me. Because I just

> ask random questions [to Google], but then, they put it together, and it's like more information that I'm actually conscious. I'm not conscious of that quantity" (D19)

4.3.2 Average Score. We calculated the sensitivity score (Fig. $5_{(2)}$) by averaging the responses to Q2 (about perceived sensitivity) for each attribute [8, 9]. Single intended interactions (i.e., training, playful, and (common) requests) are generally considered not sensitive ($sensitivity\, score \geq 1$). While single unintended interactions and the patterns and inferences derived from multiple interactions are generally considered slightly sensitive ($sensitivity\, score < 0$).

4.3.3 Spectrum. We identified a spectrum of sensitivity of speech records (Fig. 6) based on the three characteristics, *intrusiveness*, *specificity*, and *(un)availability*. Low sensitivity is attributed to individual intended interactions; where data subjects are aware data was generated and stored. These relate to simple and mundane activities and reveal little or no information about a person. This is illustrated by D6, for whom asking Google to turn on Bluetooth *"it's just an order. I didn't say anything, like personal information, nothing like that. It's just some words. It doesn't mean anything."* Additionally, they contain information that is already known or widely available (e.g., through social media, on the internet), *"that everyone can look for"* (D19). Such as someone's relationship status or city of residence. For example, D20 uses his Google Assistant to navigate the city and has multiple interactions containing information about specific stores and addresses, including work and home, *"I mean, everybody knows that. What I do, where I go shopping. It's nothing special."*

Medium sensitivity is attributed to single intended interactions that contain traces of personal information and *"tell a bit more about me"* (D18), even when these correspond to simple and mundane activities. For example, the single interaction *"what's the weather like in* [city]*?"* (medium sensitivity) *"shows a little bit about my behavior or plans"* (D2) with respect to the interaction *"what's the weather like?"* (low sensitivity). These kinds of interactions contain public personal information, but not information that is specific about a person or her private space.

High sensitivity is attributed to single intended interactions when they reveal specific information about a person's private space, including health and well-being, alcohol consumption, political opinions, interests, musical preferences, and a person's *"weird parts"* (D17), including the *"kind of stupid things I said to Google"* (D18).

> "I guess more a window into your more personal life, yeah, which I guess is by definition more sensitive." (D6)

High sensitivity is also attributed to single interactions, occurring when the speaker is unaware. This includes unintended interactions, *" Google was recording information that I was not aware of, which I would say is sensitive"* (D16), and intended interactions where the speaker is not completely aware, *"if I'm sleepy or if I'm drunk, like I'm in a state* [where] *I'm not fully aware"* (D5). These interactions often generate a sense of discomfort and awkwardness. Unawareness extends to other people (e.g., partners, family members, roommates, neighbors, visitors) tangled in the data due to its relationality. Other people's interactions, whether intended or not, are considered highly sensitive and a violation of their privacy, *"it is my uncle's privacy, I don't want to compromise, someone else's privacy"* (D2). Moreover, single interactions gain an extra layer when considering the nuances stored in the audio recordings and communicated through the voice, *"that's also something I don't think about, right? Like how do I sound at 11:33 PM when I'm asking Google to do something for me?"* (D11). Finally, high sensitivity is attributed to the patterns and inferences that can be derived from multiple interactions.

> "Even the data that you normally would say it's not sensitive, but if you put it all together it paints quite a picture of who you are and what you're doing, when, how, why." (D18)

4.4 Intimacy of Speech Records

Although we did not refer to it, the term 'intimacy' came up frequently during the interviews. Donors interpreted the intimacy of speech records in three ways. First, how speech records surface from closeness and physical proximity in a shared space. Here, intimate speech records reflect information from within that shared physical space. Second, how speech records capture intimate moments and activities. Here, speech records are derived from a device that is the viewer, or nearly participates, in intimate moments and activities. Third, how speech records capture the nuances of the daily life. Here, intimate speech records grasp what is often not material or explicit (e.g., playfulness, sleepiness, vulnerability). These three interpretations are reflected in a short excerpt where $D9_a$ explains what she means when she describes the speech records as *"very intimate"*:

> "We were listening to recordings of when we asked the alarm to turn off, and you can tell there is like a couple in bed, very sleepy, you know? And it feels like it [Google Assistant] was in bed with us, kind of, and so it's not the public voice I have when I'm online or with my friends. It's like a space that is really, really intimate, in the sense that no one else is usually allowed." ($D9_a$)

Given that the physical space is permanently shared and often corresponds to an intimate space within a person's home or close to her body, intimacy permeates through all three dimensions of speech records. Timestamp, by capturing instances of a time frame that belongs to the intimate private space, *"at 11:00 PM it's more like a private time, a more intimate time. But around the morning it's more like working* [time]*, so it doesn't feel like that intimate"* (D5); Transcript, or content, delving deeper into a person's private space by depicting her inner thoughts, ambitions, and vulnerabilities, which are often not visible to others, *"information about my thinking, my worries, and what I look for is critical. It seems very personal"* (D14); And audio recording, documenting the acoustic nuances of the space and the speaker, *"you can hear that the person doing the recording is actually sleepy, and that feels more intimate somehow"* (D2). Intimacy also derives from the interpretation and reconstruction of the data, where mundane interactions gain additional layers as they are situated in a specific context where intimate activities

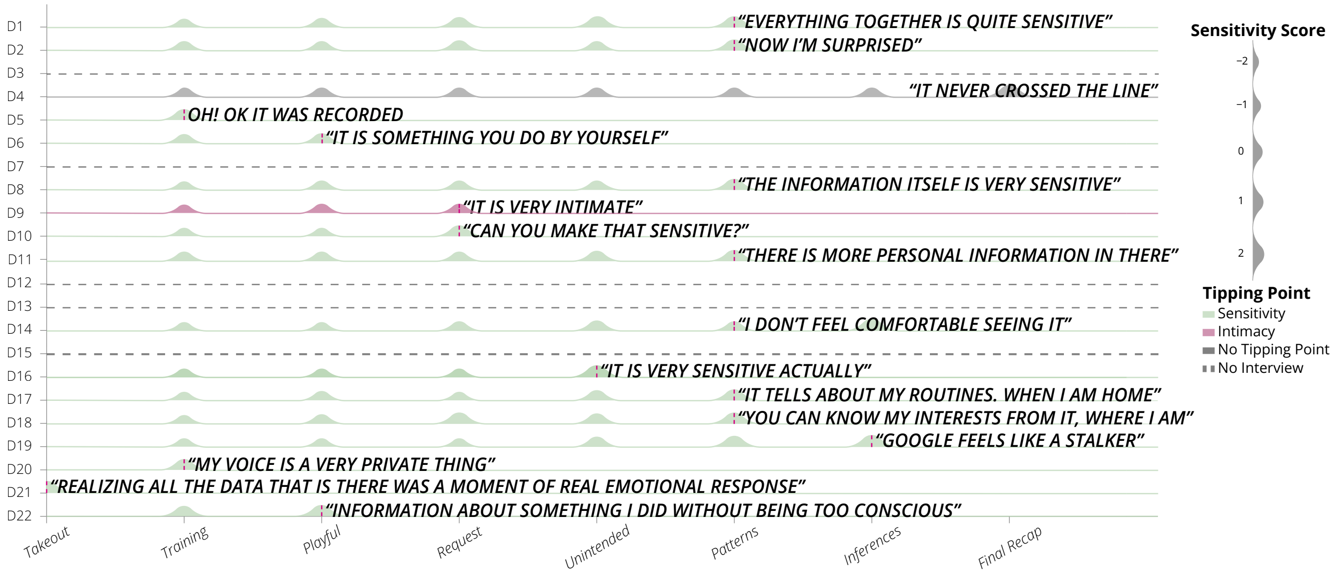


Figure 7: Tipping point in the course of data donation process and data-centric interviews.

occur. This type of intimacy is not inherent in the data but unfolds through the active participation of the data subjects.

> "The other thing is, it's not represented here [the visualization], occasionally when my partner and I have sex, we turn on and off the lights, kind of like before to arrange things and then after to clean things up. So there's maybe a detectable pattern that reveals our sex life." (D11)

Intimate data shares the characteristics of sensitive data. That is, intimate data is *intrusive* (i.e., generated through close physical proximity to a bystander device), *specific* (i.e., uniquely about a person and inherent in who they are), and *(un)available* (i.e., capturing the nuances). Yet, unlike sensitivity, intimacy is not associated with information disclosure and potential privacy violations. It relates to the information itself and its materialization. Hence, capturing and reflecting what is often not expressed or made explicit; even if mundane.

4.5 Tipping Point

Our research process involved several opportunities for donors to reflect on their speech records, including downloading a copy, exploring the takeout file, and visualizing them. Through these activities, most donors experienced a tipping point where a seemingly innocuous dataset suddenly contained sensitive and intimate data about themselves. Of the 17 donors who participated in the data-centric interviews, 16 experienced a tipping point (Figure 7). For D5, D20, and D21 this point was early and came from the paradoxical realization that speech records were collected and stored. Paradoxical since donors were aware of this, but it became even more evident by listening to individual speech records. D21 manually explored the takeout file before donating his data; for him, this realization came before the data-centric interviews. For D5 and D20 this realization came early during the data-centric interviews when listening to the first speech record (i.e., single training interaction). For D6, D9, D10, and D22, this point came from realizing that traces of private personal information or nuanced intimate information were present in individual speech records. For D16 this point came from an unintended interaction; leading to a misalignment of expectations about the device's behavior and a breach of trust. Finally, for the remaining eight donors (D1, D2, D8, D11, D14, D17, D18, D19), this point came from gaining a better understanding of the information derived from the aggregation of multiple speech records; facilitated through the personal data canvas. Namely, patterns and inferences. The tipping point underlines how sensitivity and intimacy are not self-evident but characteristics of the data that need to be disentangled and explored.

5 DISCUSSION

Our study investigated people's perspectives on sensitivity and intimacy with a focus on speech records. Our findings illustrate three characteristics of 'sensitive' and 'intimate' speech records. First, they are intrusive as they enter a person's private space, capture what happens within and expose it to unwanted access, or 'the unknown' [56]. Second, they are specific, narrowly personal (i.e., about me and no one else) and depicting a situation through time and space. Third, they are (un)available, not widespread yet potentially available. We discuss our findings (Section 5.1), discuss implications for data holders, policy makers, and researchers involved with sensitive and intimate data (i.e., speech records or the information derived from them) (Section 5.2), and provide recommendations for researchers aiming to support research participants disentangle the sensitivity and intimacy of their speech records (Section 5.3).

5.1 Understanding Sensitivity and Intimacy

Previous research on single speech records (i.e., individual interactions with a voice assistant) concludes that these are not perceived as 'sensitive' information [31, 36]. Our research partially aligns with their findings (i.e., individual interactions are generally considered not sensitive). However, it highlights the fundamental differences of comprehensively engaging with speech records. The interaction "OK Google, set an alarm", mundane and not considered sensitive in the GDPR [23], by HCI scholars [32, 34, 61, 62] (not accounting for inferences [52]), or according to previous research [31, 36], illustrates the nuances of people's perceptions of sensitive and intimate data (i.e., speech records or the information derived from them). This interaction is perceived as 'sensitive data' when it is generated from a place of playfulness (e.g., funny accent) or low consciousness (e.g., unaware, drunk). Multiple of these interactions are perceived as 'sensitive data' when information about people's behavior and routine can be inferred from them. Additionally, this interaction is perceived as 'intimate data' when it is generated from a place of vulnerability (e.g., sleepy) or it accounts for intimate contexts and activities (e.g., a couple in bed) occurring within a shared physical space.

Notions of 'sensitive data' as information that should not be disclosed [34, 61] or belonging to specific categories [23, 32, 52, 62] are limited. They do not account for the pervasive and dynamic nature of the data or potential inferences derived from aggregation. Sharing a physical space with a digital device, such as Google Assistant, means it captures various sensitive and intimate information over time, belonging to prescribed categories (e.g., political opinions) and beyond (e.g., subjectively defined private spaces). Yet, these are usually fuzzy, not evident on the surface. Sensitivity has been widely associated with private information [23, 32, 52, 62] and information disclosure [34, 61]. Therefore, it has been closely related to privacy; defining where sensitive information resides and where are the boundaries [6, 46] and selectively [44] and contextually [41] limiting disclosure. But, how can I control the disclosure of information I don't know is there? Sensitive speech records are not only a window for 'the unknown' [56] to access a person's private space but are also, to some extent, 'the unknown'.

From the three characteristics we identified, we propose to understand sensitive data as personal information that is absorbed and abstracted from a person's private space. Sensitive data is absorbed as it is generated through digital devices that intrude on a person's private physical space and transform specific instances into information that is situated and time-bounded but decontextualized. Thus, sensitive data retains information on a person's behavior that is often (un)available to her and others. It is abstracted as it detaches a person from the context and nuances of her private space. Moreover, we propose to frame 'intimate data' as a subset of 'sensitive data' that relates to the absorption and abstraction of information about intimate contexts and activities within a shared physical space. This aligns with previous literature relating intimacy to physical togetherness [27, 45] and intimate contexts and activities [4, 5, 30]. This conceptualization, and the characteristics on which it is based, rely on the distinction and boundaries between private and public information [6, 46]. Yet, these boundaries become fuzzy around data collected as a byproduct of people's interactions with a digital product or service, such as speech records. Our conceptualization of 'sensitive data' contributes to the privacy literature by underlining the need to disentangle sensitivity and intimacy from the data as a pre-requisite to defining the context-specific boundaries around it.

5.2 Implications for Data Holders, Policy Makers, and Researchers

> "I must say that I feel a bit naked. In the sense that this tells a lot about me, much more than I expected. And, in fact, it shows how technology actually is so close to us that it is able to give this amount of information."
> (D17 at the end of the interview)

Surprise and unexpectedness were common feelings among donors who experienced a tipping point (Section 4.5). Yet, they played an active role in the generation of most speech records; when giving a command or asking a question. In addition, they were informed of the data collection practices of their Google Assistant at least in two instances; when setting up their device and when they came across our research. Feelings of surprise, unexpectedness, and even nakedness illustrate how data holders (e.g., Google) and researchers (e.g., us) often fail to adequately inform their users and participants. They know that personal data is being collected, but they often do not know how it looks, how it feels, or what kind of sensitive or intimate information it contains or can be inferred from it. We discuss implications for data holders, policy makers, and researchers, aimed at mitigating surprise and unexpectedness.

- **Implications for Data Holders:** Our findings suggest that the surprise of finding unintended interactions or unraveling sensitive information from the aggregation of multiple interactions diminishes trust in the data holders. For them, mitigating surprise is an opportunity to build better relationships with their users and build trust [14]. Data holders have the opportunity to build new functionalities around personal data and provide its users with new capabilities such as data exploration and curation. Currently, Google Assistant users can review a history of their data (i.e., a scrollable list) and manually delete specific interactions. They could provide alternative data representations that empower users to engage with data as a tool to reflect and gain personal insights [14], and even actively re-define their privacy boundaries [31, 44, 60, 65].
- **Implications for Policy Makers:** The donors in our study made use of the GDPR's 'right to data portability' [23, Art. 20] to obtain a copy of their data from Google and reuse it for the purpose of contributing to our research. Although they obtained a copy of the data and could explore it autonomously, most donors found the multiple files and formats difficult to navigate. Previous research points out to similar limitations of data portability, as data is generally provided in frustrating formats and through difficult processes [3, 14]. It aligns with the purpose of 'portability' which attempts to increase end-users agency by decoupling the data from its holder. Policy makers could mitigate donors' surprise by developing policies encouraging personal data literacy [25, 63]. For example, they could require data holders to engage with

end-users to create a set of data interactions that increase their understanding of the data.

- **Implications for Researchers:** Our research is one of many that engage with personal data. During this process, we witnessed the discomfort and vulnerability that stem from surprise and the unexpected. We argue that the fuzzy and sensitive nature of personal data is an opportunity to develop a research process and agenda that integrates personal data literacy and exploration. First, it is an opportunity to facilitate the disentangling of the data through visualizations and creative sessions [22, 26, 47], and support participants in actively and iteratively defining personal boundaries. Second, it augments data with contextual information, reducing assumptions and misinterpretations for the researchers [21]. In addition, our research underlines the relational nature of personal data. It depicts people's relationships with others [48, 49], which naturally occur when personal data is generated through multi-user digital devices in public or shared spaces, such as voice assistants [19, 31, 36, 47]. Thus, accounting for others, such as third parties and minors [19], presents an open challenge, involving and informing all parties or excluding them from data collection.

5.3 Recommendations: Distangling Sensitivity and Intimacy from Speech Records

The personal data canvas was a first step in supporting users and research participants to disentangle sensitivity and intimacy from their speech records. It enabled donors to interact with their speech records in a less confusing way than the difficult-to-explore takeout file. Besides, it raised awareness about unintended interactions, patterns, and inferences. We translate our experience into practical recommendations to design processes and tools that help disentangle the sensitivity and privacy of speech records. These are directed at researchers engaging with speech records.

- **Provide a comprehensive overview:** We provided donors with multiple perspectives from which to approach their data. In contrast to previous research (e.g., [31, 36]), which focused on randomly selected single interactions, we introduced different types of single interactions and information derived through aggregation and inference. We recommend providing a comprehensive overview to progressively and incrementally approach and interrogate sensitivity and intimacy.
- **Draw attention to the unintended and the unexpected:** Donors often found sensitivity in situations of low awareness or vulnerability (e.g., unintended interactions). These are less frequent, especially in large datasets, but they are important to highlight. We recommend bringing them to the fore but acknowledging that they are rare.
- **Support structuring the data:** The personal data canvas depicted categories based on patterns of use, temporal patterns, and frequency, which donors highly appreciated. For some, the canvas was the highlight of the interview. In addition, they were helpful prompts for personal reflection and interrogation of the sensitive and intimate elements. We recommend facilitating the process of structuring the data to support interpretation and sensemaking.
- **Foster reconstructing the context of the data:** Speech records are limited and decontextualized. However, in some cases, sensitivity and intimacy derive only from the context in which they are generated. For instance, a donor recognized anomalies in her Friday routine that led to her identifying sensitive attributes in her data. We recommend creating activities that facilitate the reconstruction of the context of the data so that people understand how data relates to their day-to-day life and activities. It is an opportunity to encourage self-reflection through the data and disentangle sensitivity and intimacy.
- **Help navigate personal settings:** At the end of the interview, more than half of the donors were interested in receiving suggestions on how to configure their devices to mitigate the collection of sensitive or intimate speech records. We recommend providing general guidelines (e.g., how to disable voice data collection) and space for case-specific suggestions.

6 LIMITATIONS AND FUTURE CHALLENGES

Our research introduces the perspective of data subjects (i.e., Google Assistant users) concerning the sensitivity and intimacy of their speech records. Yet, there are limitations to our approach. First, our research is grounded and limited by the specific characteristics of speech records collected by always-on voice assistants and our choices on how to present and visualize them. Future research should explore sensitivity and intimacy in other types of data collected by digital devices (e.g., sensor data, and location) and consider alternative ways to facilitate data exploration. Second, our research involved users of Google Assistant who were willing to share their personal data with a research team. We do not incorporate the perceptions of people who refrain from using a smart assistant or are less inclined to share their personal data. They might conceive sensitivity and intimacy differently. Future research could find ways to integrate their perspective while respecting their boundaries. Third, our research was conducted with participants from a western cultural background. Future research should investigate how cultural settings influence perceptions of sensitivity and intimacy, involving a more diverse set of users. Similarly, future research should account for multi-user environments and the perspectives of other people present in the data, since data is often collected about them. Fourth, four donors had to opt-in for their Google Assistant to store their data. Although we instructed them to interact with their device as they normally would, they were aware of our research and might have altered their interaction patterns. Additionally, future research could further explore subjectively defined private spaces and how data observed by digital devices within them is perceived. Moreover, future research could investigate ways to enable and support personal data literacy that includes 'sensitive' and 'intimate' data exploration, as well as informed boundary setting.

7 CONCLUSION

In this paper, we collaborated with Google Assistant users to investigate the following research question: How do voice assistant users perceive 'sensitive' and 'intimate' data when faced with a comprehensive view of their speech records? We received speech records generated in-the-wild from 22 Google Assistant users through data donation. We analyzed the dimensions and characteristics of the received speech records in relation to previous literature and mapped them in terms of sensitivity and intimacy. We used the results of the analysis to develop a scenario-based interview protocol around data-sharing and sensitivity that we used to conduct semi-structured interviews with 17 of the 22 Google Assistant users who donated their data. We reported on our findings, suggesting that sensitivity is associated with the disclosure of information that is (1) intrusive, (2) specific, and (3) (un)available, while intimacy is a subset of sensitivity and unfolds through the recording of intimate thoughts and activities. In addition, we described the tipping point where most donors' perceptions of the sensitivity and intimacy of their speech records changed. We discussed the implications of our findings for data holders, policy makers, and researchers involved with 'sensitive' and 'intimate' data. In addition, we provided recommendations for researchers aiming to support research participants disentangle the sensitivity and intimacy of their speech records.

ACKNOWLEDGMENTS

We are deeply grateful to the people who donated their data and shared their experiences with us for their immense contributions to this research. We thank Laura and Santiago from the Cosas de Internet podcast for sharing our call with their community, Carlo van der Valk for his support with the data donation platform, and Jeff Love for his feedback. We also wish to acknowledge the valuable and thoughtful comments from the anonymous CHI'23 reviewers that helped us strengthen the paper.

REFERENCES

[1] Jacob Abbott, Haley MacLeod, Novia Nurain, Gustave Ekobe, and Sameer Patil. 2019. Local standards for anonymization practices in health, wellness, accessibility, and aging research at CHI. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, Glasgow, Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300692

[2] Noura Abdi, Xiao Zhan, Kopo M. Ramokapane, and Jose Such. 2021. Privacy Norms for Smart Home Personal Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Yokohama, Japan, 1–14. https://doi.org/10.1145/3411764.3445122

[3] Fatemeh Alizadeh, Timo Jakobi, Jens Boldt, and Gunnar Stevens. 2019. GDPR-Reality Check on the Right to Access Data. In *Proceedings of Mensch und Computer 2019*. ACM, New York, NY, USA, 811–814. https://doi.org/10.1145/3340764.3344913

[4] Teresa Almeida, Rob Comber, and Madeline Balaam. 2016. HCI and intimate care as an agenda for change in women's health. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery (ACM), San Jose, CA, USA, 2599–2611. https://doi.org/10.1145/2858036.2858187

[5] Teresa Almeida, Rob Comber, Gavin Wood, Dean Saraf, and Madeline Balaam. 2016. On looking at the vagina through Labella. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery (ACM), San Jose, CA, USA, 1810–1821. https://doi.org/10.1145/2858036.2858119

[6] Irwin Altman. 1975. The Environment and Social Behaviour: Privacy, Personal Space, Territory and Crowding. ERIC.

[7] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive self-disclosures, responses, and social support on instagram: The case of #depression. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. Association for Computing Machinery, Portland, Oregon, USA, 1485–1500. https://doi.org/10.1145/2998181.2998243

[8] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering smart home internet of things privacy norms using contextual integrity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2 (2018), 23. https://doi.org/10.1145/3214262

[9] Noah Apthorpe, Sarah Varghese, and Nick Feamster. 2019. Evaluating the contextual integrity of privacy regulation: Parents' IoT toy privacy norms versus COPPA. In *Proceedings of the 28th USENIX Conference on Security Symposium*. USENIX Association, Santa Clara, CA, 123–140. https://dl.acm.org/doi/abs/10.5555/3361338.3361348

[10] Madeline Balaam, Rob Comber, Ed Jenkins, Selina Sutton, and Andrew Garbett. 2015. Feedfinder: A location-mapping mobile application for breastfeeding women. *Conference on Human Factors in Computing Systems - Proceedings* 2015-April (2015), 1709–1718. https://doi.org/10.1145/2702123.2702328

[11] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (9 2018), 1–24. https://doi.org/10.1145/3264901

[12] Sander Bogers, Joep Frens, Janne van Kollenburg, Eva Deckers, and Caroline Hummels. 2016. Connected Baby Bottle. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. ACM, New York, NY, USA, 301–311. https://doi.org/10.1145/2901790.2901855

[13] Jacky Bourgeois, Janet van der Linden, Gerd Kortuem, Blaine A. Price, and Christopher Rimmer. 2014. Conversations with my washing machine. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, New York, NY, USA, 459–470. https://doi.org/10.1145/2632048.2632106

[14] Alex Bowyer, Jack Holt, Josephine Go Jefferies, Rob Wilson, David Kirk, and Jan David Smeddinck. 2022. Human-GDPR Interaction: Practical Experiences of Accessing Personal Data. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3501947

[15] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (1 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[16] Virginia Braun and Victoria Clarke. 2013. *Successful Qualitative Research. A Practical Guide for Beginners*. SAGE Publications Ltd, London. 400 pages.

[17] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR FORUM* 36, 2 (2002), 3–10. https://doi.org/10.1145/792550.792552

[18] George Chalhoub, Martin J. Kraemer, Norbert Nthala, and Ivan Flechais. 2021. "It did not give me an option to decline": A Longitudinal Analysis of the User Experience of Security and Privacy in Smart Home Products. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445691

[19] Peng Cheng and Utz Roedig. 2022. Personal Voice Assistant Security and Privacy - A Survey. *Proc. IEEE* 110, 4 (2022), 476–507. https://doi.org/10.1109/JPROC.2022.3153167

[20] Andy Crabtree, Peter Tolmie, and Will Knight. 2017. Repacking 'Privacy' for a Networked World. *Computer Supported Cooperative Work: CSCW: An International Journal* 26, 4-6 (2017), 453–488. https://doi.org/10.1007/s10606-017-9276-y

[21] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. MIT Press. 328 pages. https://doi.org/10.7551/mitpress/11805.001.0001

[22] Sandra Gabriele and Sonia Chiasson. 2020. Understanding Fitness Tracker Users' Security and Privacy Knowledge, Attitudes and Behaviours. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376651

[23] GDPR. 2018. General Data Protection Regulation. , 88 pages. https://gdpr.eu/

[24] Alejandra Gómez Ortega, Jacky Bourgeois, and Gerd Kortuem. 2022. Reconstructing Intimate Contexts through Data Donation: A Case Study in Menstrual Tracking Technologies. In *Nordic Human-Computer Interaction Conference*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3546155.3546646

[25] Jonathan Gray, Carolin Gerlitz, and Liliana Bounegru. 2018. Data infrastructure literacy. *Big Data and Society* 5, 2 (2018). https://doi.org/10.1177/2053951718786316

[26] Bastian Greshake Tzovaras, Misha Angrist, Kevin Arvai, Mairi Dulaney, Vero Estrada-Galiñanes, Beau Gunderson, Tim Head, Dana Lewis, Oded Nov, Orit Shaer, Athina Tzovara, Jason Bobe, and Mad Price Ball. 2019. Open Humans: A platform for participant-centered research and personal data exploration. *GigaScience* 8, 6 (2019), 1–13. https://doi.org/10.1093/gigascience/giz076

[27] Konstantinos Grivas. 2006. Digital Selves: Devices for intimate communications between homes. *Personal and Ubiquitous Computing* 10, 2-3 (2006), 66–76. https://doi.org/10.1007/s00779-005-0003-1

[28] Karey Helms. 2019. Do you have to pee? A design space for intimate and somatic data. In *DIS 2019 - Proceedings of the 2019 ACM Designing Interactive Systems Conference*. Association for Computing Machinery (ACM), San Diego, CA, USA, 1209–1222. https://doi.org/10.1145/3322276.3322290

[29] Albrecht Kurze, Andreas Bischof, Sören Totzauer, Michael Storz, Maximilian Eibl, Margot Brereton, and Arne Berger. 2020. Guess the Data: Data Work to Understand How People Make Sense of and Use Simple Sensor Data from Homes. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376273

[30] Hyosun Kwon, Joel E. Fischer, Martin Flintham, and James Colley. 2018. The Connected Shower. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (12 2018), 1–22. https://doi.org/10.1145/3287054

[31] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening? *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (11

2018), 1–31. https://doi.org/10.1145/3274371

[32] Hong Li, Pradthana Jarusriboonchai, Heiko Müller, Emmi Harjuniemi, and Jonna Häkkilä. 2020. Emotional Communication between Remote Couples: Exploring the Design of Wearable Ambient Displays. , 5 pages. https://doi.org/10.1145/3419249.3420139

[33] Yifang Li, Nishant Vishwamitra, Hongxin Hu, and Kelly Caine. 2020. Towards A Taxonomy of Content Sensitivity and Sharing Preferences for Photos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Honolulu, HI, USA, 1–14. https://doi.org/10.1145/3313831.3376498

[34] Yin Liu, Zheng Song, and Eli Tilevich. 2017. Querying invisible objects: Supporting data-driven, privacy-preserving distributed applications. In *Proceedings of the 14th International Conference on Managed Languages and Runtimes*. Association for Computing Machinery, Prague, Czech Republic, 60–72. https://doi.org/10.1145/3132190.3132206

[35] Daniel M. Low, Kate H. Bentley, and Satrajit S. Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology* 5, 1 (2 2020), 96–116. https://doi.org/10.1002/lio2.354

[36] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. 2019. Privacy Attitudes of Smart Speaker Users. *Proceedings on Privacy Enhancing Technologies* 2019, 4 (10 2019), 250–271. https://doi.org/10.2478/popets-2019-0068

[37] Lydia Manikonda, Aditya Deotale, and Subbarao Kambhampati. 2018. What's up with Privacy? User Preferences and Privacy Concerns in Intelligent Personal Assistants. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New Orleans, LA, USA, 229–235. https://doi.org/10.1145/3278721.3278773

[38] Kirsten Martin and Helen Nissenbaum. 2015. Measuring Privacy: An Empirical Test Using Context to Expose Confounding Variables. *Columbia Science and Technology Law Review* (2015), 1–40. https://doi.org/10.2139/ssrn.2709584

[39] Maryam Mehrnezhad and Teresa Almeida. 2021. Caring for intimate data in fertility technologies. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery (ACM), Yokohama, Japan, 11. https://doi.org/10.1145/3411764.3445132

[40] Nicole Meng, Dilara Keküllüoğlu, and Kami Vaniea. 2021. Owning and Sharing. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–29. https://doi.org/10.1145/3449119

[41] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79, 1 (2004), 101–139.

[42] Chien Shing Ooi, Kah Phooi Seng, Li Minn Ang, and Li Wern Chew. 2014. A new approach of audio emotion recognition. *Expert Systems with Applications* 41, 13 (2014), 5858–5869. https://doi.org/10.1016/j.eswa.2014.03.026

[43] Antti Oulasvirta, Aurora Pihlajamaa, Jukka Perkiö, Debarshi Ray, Taneli Vähäkangas, Tero Hasu, Niklas Vainio, and Petri Myllymäki. 2012. Long-term effects of ubiquitous surveillance in the home. In *UbiComp'12 - Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. Association for Computing Machinery, 41–50. https://doi.org/10.1145/2370216.2370224

[44] Leysia Palen and Paul Dourish. 2003. Unpacking "Privacy" for a Networked World. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Ft. Lauderdale, Florida, USA, 129–136. https://doi.org/10.1145/642611.642635

[45] Emmi Parviainen and Marie Louise Juul Søndergaard. 2020. Experiential Qualities of Whispering with Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376187

[46] Sandra Petronio. 2002. *Boundaries of privacy: Dialectics of disclosure*. State University of New York Press, Albany.

[47] Dominik Pins, Timo Jakobi, Alexander Boden, Fatemeh Alizadeh, and Volker Wulf. 2021. Alexa, We Need to Talk: A Data Literacy Approach on Voice Assistants. In *Designing Interactive Systems Conference 2021*. ACM, New York, NY, USA, 495–507. https://doi.org/10.1145/3461778.3462001

[48] Barbara Prainsack. 2019. Data Donation: How to Resist the iLeviathan. In *Philosophical Studies Series*. Vol. 137. 9–22. https://doi.org/10.1007/978-3-030-04363-6{_}2

[49] Barbara Prainsack. 2019. Logged out: Ownership, exclusion and public value in the digital data and information commons. *Big Data & Society* 6, 1 (1 2019), 205395171982977. https://doi.org/10.1177/2053951719829773

[50] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. Association for Computing Machinery, Ann Arbor, MI, USA, 989–992. https://doi.org/10.1145/3209978.3210124

[51] Tauhidur Rahman, Alexander T. Adams, Mi Zhang, Erin Cherry, Bobby Zhou, Huaishu Peng, and Tanzeem Choudhury. 2014. BodyBeat: A mobile system for sensing non-speech body sounds. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*. Association for Computing Machinery, Bretton Woods, New Hampshire, USA, 2–13. https://doi.org/10.1145/2594368.2594386

[52] Anna Rudnicka, Anna L. Cox, and Sandy J. J. Gould. 2019. Why Do You Need This?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300622

[53] Koustuv Saha, Jordyn Seybolt, and Stephen M. Mattingly. 2021. What life events are disclosed on social media, how, when, and by whom?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Yokohama, Japan, 22. https://doi.org/10.1145/3411764.3445405

[54] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?". In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[55] Chen Shani, Alexander Libov, Sofia Tolmach, and Liane Lewin-eytan. 2022. " Alexa , Do You Want to Build a Snowman ?" Characterizing Playful Requests to Conversational Agents. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts*. Association for Computing Machinery (ACM), New Orleans, LA, USA, 1–7. https://doi.org/10.1145/3491101.3519870

[56] Irina Shklovski, Scott D. Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. 2014. Leakiness and creepiness in app space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Toronto, Ontario, Canada, 2347–2356. https://doi.org/10.1145/2556288.2557421

[57] Yan Shvartzshnaider, Thomas Wies, Paula Kift, and Helen Nissenbaum. 2016. Learning Privacy Expectations by Crowdsourcing Contextual Informational Norms. *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing* 4, 1 (2016), 209–218. https://ojs.aaai.org/index.php/HCOMP/article/view/13271

[58] Anya Skatova and James Goulding. 2019. Psychology of personal data donation. *PLOS ONE* 14, 11 (11 2019), e0224240. https://doi.org/10.1371/journal.pone.0224240

[59] Sara Skoog Waller, Mårten Eriksson, and Patrik Sörqvist. 2015. Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology* 6, July (2015), 1–11. https://doi.org/10.3389/fpsyg.2015.00978

[60] Veronika Strotbaum, Monika Pobiruchin, Björn Schreiweis, Martin Wiesner, and Brigitte Strahwald. 2019. Your data is gold – Data donation for better healthcare? *it - Information Technology* 61, 5-6 (10 2019), 219–229. https://doi.org/10.1515/itit-2019-0024

[61] Vasileios Tsoukas, Anargyros Gkogkidis, and Athanasios Kakarountas. 2020. A Survey on Mobile User Perceptions of Sensitive Data and Authentication Methods. In *24th Pan-Hellenic Conference on Informatics*. Association for Computing Machinery, Athens, Greece, 346–349. https://doi.org/10.1145/3437120.3437337

[62] Mark Warner, Agnieszka Kitkowska, Jo Gibbs, Juan F. Maestre, and Ann Blandford. 2020. Evaluating 'Prefer not to say' Around Sensitive Disclosures. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376150

[63] Annika Wolff, Daniel Gooch, Jose Cavero Montaner, Umar Rashid, and Gerd Kortuem. 2017. Creating an understanding of data literacy for a data-driven society. *Journal of Community Informatics* 12, 3 (2017), (In press). www.ci-journal.net/index.php/ciej/article/view/1286.

[64] World Economic Forum. 2011. *Personal Data : The Emergence of a New Asset Class An Initiative of the World Economic Forum*.

[65] Matthew Zook, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A. Koenig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson, and Frank Pasquale. 2017. Ten simple rules for responsible big data research. *PLoS Computational Biology* 13, 3 (2017), 1–10. https://doi.org/10.1371/journal.pcbi.1005399