Int. J. Human–Computer Studies 205 (2025) 103590

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Exploring the Impact of Modality and Speech Rate Manipulation in Voice Permission Requests—Limits of Applicability and Potential for Influencing Decision-Making

Anna Leschanowsky [a,1,*], Anastasia Sergeeva [b,1], Judith Bauer [a], Sheetal Vijapurapu [a], Mateusz Dubiel [b]

[a] Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
[b] University of Luxembourg, Esch-sur-Alzette, Luxembourg

ARTICLE INFO

Keywords:
Auditory feedback
Synthetic speech
Design ethics

ABSTRACT

As voice-enabled technologies are becoming increasingly more prevalent, voice-enabled permission requests become a crucial topic of investigation. It is yet unclear how to appropriately inform users in voice user interfaces (VUIs) about data processing practices. To understand how modality (text vs. voice) and the speech rate of the voice can influence users' perceptions and decisions to grant permission, we conducted two preregistered studies (N = 343 and N = 594) and one pre-study, including two listening tasks to design potentially deceptive voice patterns. We found that users can distinguish between different levels of intrusiveness in the voice modality. However, they are less likely to accept voice-based permissions, pointing to cognitive problems associated with them. Moreover, we found that speech rate manipulations of action verbs "Accept" and "Decline" shifted users' decisions towards acceptance, making the effect less controllable than predicted. This work highlights implications and design considerations for future voice-enabled permission requests.

1. Introduction

Voice-enabled conversational assistants, such as Amazon Alexa, Apple Siri, or Google Home are becoming increasingly more popular. According to the Voice Consumer Index 2023, voice assistant usage ranges from 49% in Germany to over 60% in the US and UK and up to 66% in Mexico (Vixen Labs, 2023). While voice assistants are still predominantly used for simple tasks such as searching the web, checking the weather, playing music, or setting alarms (Ammari et al., 2019; Vixen Labs, 2023), there is a growing trend towards using these devices routinely for purchasing products and services online (National Public Media, 2022).

To expand the range of services offered by voice assistants, platforms such as Amazon Alexa or Google Assistant allow third-party developers to create and publish voice applications. As of today, over 130.000 Alexa Skills are available, compared to just 135 in early 2016 (Alexa Developer, 2023; Edu et al., 2021). Once published, users can enable voice applications either by using an invocation command or through the platform's application store. Just like traditional mobile apps, voice applications must comply with legal regulations by providing privacy policies and disclosing data practices. To ensure compliance and protect users, platforms have implemented policy requirements and skill certification processes.[2] However, despite these measures, previous research has shown that voice applications can circumvent certification processes and violate privacy policies (Liao et al., 2020; Cheng et al., 2020; Liao et al., 2023). Despite their importance, privacy policies are not accessible within the conversational interface, leaving it to the users to read them (Liao et al., 2020). However, if personal information is collected, a permission request is sent to the voice assistants' companion app for user approval. To improve the usability of these requests, platform providers have started integrating them directly into the voice interface.[3]

* Corresponding author.
E-mail address: anna.leschanowsky@iis.fraunhofer.de (A. Leschanowsky).
[1] Contributed equally to the manuscript.
[2] https://developer.amazon.com/en-US/docs/alexa/custom-skills/certification-requirements-for-custom-skills.html.
[3] https://developer.amazon.com/en-US/docs/alexa/custom-skills/use-voice-forward-consent.html.
[4] https://developer.amazon.com/en-US/docs/alexa/custom-skills/configure-permissions-for-customer-information-in-your-skill.html.
[5] https://developer.android.com/guide/topics/permissions/overview.

https://doi.org/10.1016/j.ijhcs.2025.103590
Received 15 January 2025; Received in revised form 21 July 2025; Accepted 28 July 2025
Available online 10 September 2025

To collect user data, developers can request information directly through the platform's API. For instance, Alexa Skills can request access to customer names, email addresses, as well as location data.[4] Similarly, Google provides a more structured permission system by categorizing requests by type and sensitivity with certain permissions being labeled "dangerous" and requiring explicit consent.[5] Restricted permission requests can access personal and sensitive user data such as location, photos or all files. While platforms emphasize the importance of data minimization, control and transparency, previous research has highlighted unjustified data collection in voice applications (Edu et al., 2021). Edu et al. (2022) assessed Alexa Skill privacy practices over three years and found that while the majority of voice applications did not request any permissions, the number of permission requests increased over time. Additionally, voice applications asking for sensitive information directly through the conversational interface, bypassing the formal API permission process, continue to raise privacy concerns (Lentzsch et al., 2021; Edu et al., 2022).

Although permission requests play a critical role in protecting user privacy, they open up ways for implementing so-called "deceptive/dark[6] design patterns" – instances of the interface designs that can potentially affect users' decisions towards their disadvantage (Gray et al., 2018). The European Union has acknowledged the manipulative potential of deceptive patterns and has addressed them in various guidelines and regulations such as the EDPB Guidelines on Dark patterns in social media platform interfaces, the Digital Service and Markets Act and the AI Act (European Data Protection Board (EDPB), 2023; European Parliament and Council of the European Union, 2022a,b, 2024). Notably, Article 5 of the AI Act prohibits AI systems that *"deploy purposefully manipulative or deceptive techniques with the objective, or the effect of materially distorting the behavior of a person [...] thereby causing them to take a decision that they would not have otherwise taken [...]"*. However, while these regulations focus on deceptive patterns in general or on visual interfaces in particular, the issue of deceptive patterns in voice-based applications remains largely overlooked in legal frameworks (De Conca, 2023). While the potential of dark patterns to influence users' interactions with voice-based systems has been recently highlighted (Owens et al., 2022), this area of research remains understudied. As permission requests shift from the screen to the conversational interface, it becomes crucial to understand users' perceptions of voice-based permission requests and to identify potential deceptive practices. In this work, we investigate how users perceive voice-based permission requests in comparison to text-based requests and to what extent they may be susceptible to manipulation.

In sum, through this work:

- We provide quantitative and qualitative insights into users' perception of voice-based permission requests compared to text-based requests under controlled conditions.
- We contribute to the methodology for investigating modality-specific features in the manipulation of user choice, using speech rate manipulation as an example.

2. Background and related work

Below, we summarize four bodies of work relevant to our studies by reviewing the literature on user privacy and consent in voice user interfaces (VUIs), research on deceptive patterns in VUIs, the effect of modality on perception and interaction with information and relevant prosodic parameters that may affect the perception of speech and user behavior.

2.1. Privacy, consent and permission requests in voice-based technology

With the growing adoption of voice-based technology, concerns around privacy and security are rising in parallel (Fruchter and Liccardi, 2018; Lau et al., 2018; Seymour et al., 2020). Although privacy lacks a universally accepted definition, it is often conceptualized in terms of control (Altman, 1975; Westin, 1967; Smith et al., 2011). In the context of conversational AI, privacy as control is usually understood as users' ability to control existing personal information as well as personal data generated in the future, either related to personal data disclosure or data access and retention (Leschanowsky et al., 2024; Kang and Oh, 2023; Pal et al., 2020; Lin and Parkin, 2020). The aspect of control is also emphasized by legal regulations around the world, such as the General Data Protection Regulation (GDPR), which provides clear guidelines on consent and user rights when processing personal data (European Commission, 2016). Moreover, offering users control options within voice-based interfaces can strengthen user trust and empowerment (Ahmad et al., 2022).

Privacy information and privacy controls can take various forms, including consent mechanisms, privacy notices, labels and settings. Previous research explored strategies to appropriately inform users in a VUI (Harkous et al., 2016) and developed conversational privacy strategies to enhance user control (Brüggemeier and Lalone, 2022; Leschanowsky et al., 2023). However, consent and permission requests in VUIs are still frequently carried out on screen-based technologies. For instance, previous research investigated runtime permission requests in proactive assistants using a screen accompanied by an audible bell (Malkin et al., 2022). Yet, requiring users to switch modalities can be a barrier to meaningful consent, as it can increase cognitive load and can lead to user errors (Sandhu and Dyson, 2012).

To improve the usability of consent requests in voice-based technologies, platform providers have integrated consent and permission requests directly into the VUI. For instance, Alexa's "Voice Forward Consent" feature (VFC)[7] allows developers to integrate verbal permission requests that enable users to approve or deny requests via voice. In the European Union, consent mechanisms are an essential part of regulation when placing or reading information from users' devices under the ePrivacy Directive or when processing personal data under the GDPR (Data Protection Working Party, 2013). Thus, verbal consent mechanisms are essential for both legal compliance and safeguarding user privacy in voice-based interfaces. However, they also introduce significant challenges, such as pressure (i.e., perceived urgency to act) and the limited amount of information that can be delivered through VUIs (Seymour et al., 2022). Recent research has begun to develop recommendations for regulators, voice assistant developers and platforms to improve the design of verbal consent mechanisms (Seymour et al., 2023, 2024). While research on verbal consent and permission request mechanisms is still limited, privacy notice research in GUIs is an active field, and its design recommendations can inform research and implementation in the voice domain (Schaub et al., 2015; Utz et al., 2019).

2.2. Deceptive design patterns in VUI

In the domain of voice-based technology, even beyond the consent and permission problem, there are significant concerns about the ethics of VUI interaction designs. For example, an expert-based study conducted by Mildner et al. (2024) identified areas of unethical and privacy-invasive designs which can easily be implemented in the voice interaction domain. These areas include the lack of privacy protection mechanisms and the potential for deceptive interactions. While unethical design solutions may sometimes arise from short-sighted design

[6] We acknowledge the current shift in the community from the term "dark patterns" to "deceptive patterns". In this paper, we will use the terms interchangeably and refer to the names of existing regulations as they were originally established.

[7] https://developer.amazon.com/en-US/docs/alexa/custom-skills/use-voice-forward-consent.html.

choices (Gray et al., 2020) or difficulties in translating visual-based interaction parameters into the voice domain, previous work discusses the possibility of deliberately distorting user choices in the interests of voice application providers (Owens et al., 2022; Mildner et al., 2024). Here, nudges and deceptive design patterns, already well explored in visual and text-based applications (e.g., Bongard-Blanchy et al. (2021), Tuncer et al. (2023) and Gray et al. (2021)), can also increase the manipulative potential of VUIs.

Deceptive design patterns are usually described as user interface elements, which can push users' decisions in specific directions, such as choosing a more profitable subscription option for the company or sharing extensive personal data (Mathur et al., 2021). They are purposely designed to confuse users, hinder them from expressing their true preferences, or coerce them into specific actions (Gray et al., 2018). Studies on dark patterns in the visual domain have shown that users more easily recognize some practices (e.g., fake urgency) than others (e.g., forced consent) (Bongard-Blanchy et al., 2021), and that more covert, subtle manipulations often yield better results for companies (Luguri and Strahilevitz, 2021). While multiple studies were conducted in the field of visual interfaces, the area of voice-based interaction is still underexplored. Owens et al. (2022) sought expert opinions on potential dark patterns in VUIs, focusing on a range of theoretical scenarios, including both interaction parameters of voice assistant technology and speech properties such as volume, pitch, rate, fluency, pronunciation, and articulation to emphasize certain options and, consequently, increase their prominence to the user. Dula et al. (2023) discussed the parameters of the voice as part of "dishonest anthropomorphism" a deceptive design feature whereby the human-likeness of the agent is used to influence users.

While deceptive practices can affect users' choices in many ways, particular attention is paid to the link between deceptive patterns and the infringement of users' privacy (Bösch et al., 2016; Gunawan et al., 2022; Valoggia et al., 2024). Several studies addressed the problem of cookie banners (Krisam et al., 2021; Gray et al., 2021) and the ways companies can use the asymmetric choice presentation to make privacy-unfriendly options more salient, thereby pushing users to select them without consciously processing the options.

In the area of VUIs, similar ways of presenting the asymmetric choice were explored (Dubiel et al., 2024b). The results revealed a significant impact of synthetic voice fidelity (i.e., the degree of resemblance of natural speech) on people's decision-making and that these effects often go unnoticed by users. However, the findings are primarily relevant to multi-agent interaction scenarios and focused on general decision-making rather than privacy-related choices (Dubiel et al., 2024b).

2.3. Effect of the modality to the perception of the information

Various studies discussed the difference between text and speech modalities in the process of information acquisition. In an educational assessment setting, the information presented via voice is perceived as having a higher quality compared to text (Wambsganss et al., 2022). In the information search task, the interaction with a speech-based agent showed higher perceived efficiency, lower cognitive effort, higher enjoyment, and higher service satisfaction than a text-based agent (Rzepka et al., 2022). At the same time, in the domain of customer service, voice interaction was perceived as having higher cognitive demands compared to text-based ones and, therefore, activated critical thinking and diminished persuasive effects (Ischen et al., 2022). However, compared to text messages, voice messages are perceived as more credible, even if they have problems with accuracy and information attribution (Gaiser and Utz, 2023). In the presentation of the information about new technology, voice modality had a significant positive effect on the attitude towards the described technology (Geipel et al., 2023).

2.4. Speech prosody and speech perception

Social impressions of others are significantly shaped by their voices, particularly through non-verbal cues like intonation, emphasis, and rhythm (Belin et al., 2011). These cues influence both how we perceive a speaker (Varghese and Nilsen, 2020) and process information (Rodero, 2016). Among the various prosodic aspects of the human voice, mean fundamental frequency (F0), judged as voice pitch (Titze and Martin, 1998) and speech rate (Dowding et al., 2024) are central for making social judgments (Schild et al., 2020).

Speech rate is broadly defined in the phonetics literature as the number of spoken units per unit time (e.g., Tsao et al. (2006)). Speech rate provides a global measure of the pace at which a speaker constructs and produces speech, and includes broader characteristics, such as a tendency to hesitate while contemplating the next utterance (Dowding et al., 2024). Faster speakers tend to be rated as more convincing, reliable, empathic, serious, active, and competent (Apple et al., 1979). Younger speakers are generally perceived as more trustworthy (Ernst and Herm-Stapelberg, 2020; Pias et al., 2024), and a higher pitch can enhance credibility, while faster speech is more persuasive than slower speech (Schirmer et al., 2020). However, it should be noted that preferences may depend on the application domain (Goodman and Mayhorn, 2023) and differ across cultures and user demographics (Bem, 1981). Additionally, a moderate speech pace of around 180 words per minute is considered optimal for information recall and recognition (Rodero, 2016). Overall, speech should be fast enough to engage listeners but not so fast as to hinder comprehension (Rodero et al., 2022).

While research consistently links fast-to-moderate speech and lower pitch to higher perceived attractiveness and dominance in men (Puts et al., 2016; Oleszkiewicz et al., 2017; Vukovic et al., 2011), the impact of voice perception on user behavior in decision-making tasks remains understudied (Dubiel et al., 2020). In this work, we focus on the exploration of the impact of systematic speech rate manipulations on the perception of privacy permissions and users' likelihood to accept them.

Kochanski et al. (2005) found that loudness and duration patterns are primarily used by speakers to mark prominence in stress-timed languages such as English, German or Russian. This goes against textbooks (e.g., Roca and Johnson (1999) and Clark and Yallop (1996)) and common assumptions that highlight the main role of the fundamental frequency (f0) in distinguishing prominent syllables from the rest of the utterance. Previous studies also indicate that, in contrast to pitch accenting, speech rate can function as a prosodic prime, the same way as other primers (Tooley et al., 2018). Priming generally refers to the facilitative effect of an event or action on subsequent associated responses (Molden, 2014). For example, people often prefer the stimulus that was associated with a prime presented earlier in a choice task. Studies showed that the speech rate of the primer affected the speech rate of the post-priming production (Jungers and Hupp, 2009; Tooley et al., 2018). While no prior studies address specifically auditory primes in decision-making, previous research in other areas (e.g. priming via visual stimuli) has shown that presenting primes beyond focused attention can influence choices in free-choice scenarios, particularly under uncertainty (Payne et al., 2016; Kiesel et al., 2006).

In speech synthesis systems, modifications of pitch, volume, and speech rate can be utilized to create models that allow for controllable emphasis (i.e., the option to deliberately change the emphasis of selected words during synthesis). This can be achieved by adjusting e.g., the pitch contour or phoneme durations (Raitio et al., 2022; Shechtman et al., 2021). Another approach is to predict features such as pitch variance, phoneme duration variance, or wavelet based features, which serve as latent emphasis scores and steer the pitch, energy, and duration of the synthesized speech (Seshadri et al., 2022). Joly et al. (2023) show that increasing phoneme durations, thereby slowing down the speech rate for a certain word, can be sufficient to produce emphasized words.

Few user interfaces allow users to control the system's speech rate directly. For example, Siri's accessibility settings include a "speaking rate" slider, visually represented by a slow-moving turtle and a fast-moving hare (Dowding et al., 2024).

3. Research gaps and the summary of the research

Based on the relevant literature discussed in Section 2, we have identified the following gaps:

1. Previous studies have highlighted the importance of presenting privacy information in various formats and modalities (Morel and Pardo, 2020; Harkous et al., 2016), which supports the rationale for presenting consent and permission requests in the voice modality. However, to the best of our knowledge, no studies have investigated how a change in modality affects users' perception of consent and their willingness to accept permission requests.
2. Although some Deceptive Design patterns may be similar across both voice and screen-based modalities, since they operate at the information level (e.g., "Nagging" Gray et al., 2018), there are also manipulations that work at the modality level (e.g., "Preselection" for visual-based interfaces). While the potential use of modality-based patterns has been mentioned in the literature (Owens et al., 2022), only a few studies have examined this effect (Dubiel et al., 2024b). Specifically, no studies have focused on the influence of modality in a single-agent choice representation.

To fill these gaps we conducted two preregistered experimental studies (referred to as Study 1 and Study 2 in the following sections).

The research sections are organized as follows:

Study 1 (presented in Section 4) focuses on understanding the differences between perceptions of voice- and text-based permission requests, particularly in the context of varying levels of request intrusiveness. To do so, we generated a set of permission requests based on previous literature and tested them across two modalities (text and voice-based) and different levels of intrusiveness.

Study 2 (presented in Section 5) is based on the qualitative and quantitative results of Study 1 , highlighting differences in perception and acceptance of voice-based requests. We test whether manipulating voice-specific features affects users' willingness to accept the permission request. To do so, we investigate the effect of speech rate in the performative clause of the request (proposing either "Accepting" or "Declining" the permission) on user behavior. We select speech rate as the feature to manipulate, as previous literature has shown its priming and persuasive effects in a wide range of scenarios (Schirmer et al., 2020; Chattopadhyay et al., 2003). The pre-study (Section 5.1) outlines the considerations and strategies for creating the voice samples for Study 2, while Study 2 (Section 5) presents the experimental procedure and results of applying this manipulation strategy in the experiment. The full procedure is illustrated in Fig. 1.

4. Study 1: Effects of text- and voice-based permission requests on user perception and decision-making

In Study 1, we explore how shifting from screen-based to voice-based permission requests affects users' decision-making and perceptions. While screen-based requests can disrupt user experience due to modality switching, the impact of voice-based permission requests on user perception remains unclear. Consequently, Study 1 investigates the users' perception of permission requests. Specifically, we investigate how the modality of permission requests influences the likelihood of users accepting them and how varying levels of intrusiveness may influence this. Additionally, we assess how factors like clarity and comprehension, which are considered crucial for users' privacy decision-making (Masotina and Spagnolli, 2022), are affected by the modality of the permission request. By understanding the implications of modalities, we aim to assess benefits and challenges of voice-based permission requests.

We formulated the following research questions:

Research Question 1 (RQ1): How do users perceive differences in app permission intrusiveness, and how does their willingness to accept the permission vary when requests are presented via voice versus text across varying levels of intrusiveness? More specifically, we hypothesized that acceptance rates for app permissions will decrease as the level of data intrusiveness increases for both voice and text modalities **(H1)** and that participants will rate the intrusiveness of app permissions higher in scenarios where the actual intrusiveness of the app permissions is higher **(H2)**.

Research Question 2 (RQ2): Are there significant differences in users' acceptance of app permissions with varying levels of intrusiveness when delivered through voice compared to text modalities? The results of previous studies showed that the effect could lie in both directions (Ischen et al., 2022; Rzepka et al., 2022). Depending on the tasks, people can prefer voice-based interactions because of ease of use and lower cognitive effort (Rzepka et al., 2022), or find it more difficult to understand than text-only, influencing persuasiveness (Berry et al., 2005; Ischen et al., 2022). Therefore, we hypothesize that the acceptance rate of voice-based permissions could be higher **(H3a)** or lower **(H3b)** compared to the text-based permission requests. In the same way, we hypothesized that the voice-based permissions could be rated as less intrusive **(H4a)** or more intrusive **(H4b)** compared to text-based permissions.

Research Question 3 (RQ3): Does the modality of a permission request (voice vs. text) affect the person's impression of the clarity and understandability of the request? Again, as previous studies pointed to the bi-directional effect of the modality on the clarity and understandability of information, we formulate the hypothesis as follows: The perception of clarity and comprehensibility will be higher **(H5a)** or lower **(H5b)** in the text-based permission request than in the voice-based request.

4.1. Test materials

4.1.1. Permission requests

We designed permission requests for three distinct voice applications: a ride-hailing application "Ride Hailer" inspired by an Alexa skill,[8] a fitness application "Fit Buddy" and a recipe application "Kitchen Wizard". We did not aim to create legally valid verbal consent requests under the GDPR, as previous studies are inconclusive on what information should be included (Seymour et al., 2023), and we intentionally increased intrusiveness to assess its impact on user decision-making. Moreover, we did not address a potential conflation between permission requests and legal consent, particularly since some of our applications might access personal data under legal bases of legitimate interest or contract rather than consent. Instead, we aimed to create verbal permission requests that include information that could potentially be found in verbal consent, such as the personal data requested, the purpose of data collection, data sharing practices and data retention periods (Seymour et al., 2023). We did not include all potentially required information, such as user rights or instructions on how to withdraw consent. Nevertheless, our permission requests are the first attempt to create voice-based permission requests, including information beyond requested data types, while keeping it short and specific (Seymour et al., 2022; Schaub et al., 2017).

We designed three levels of intrusiveness, i.e., low, medium, and high intrusion (illustrated in Table 1) by varying the information that

[8] https://developer.amazon.com/en-US/docs/alexa/custom-skills/use-voice-forward-consent.html.

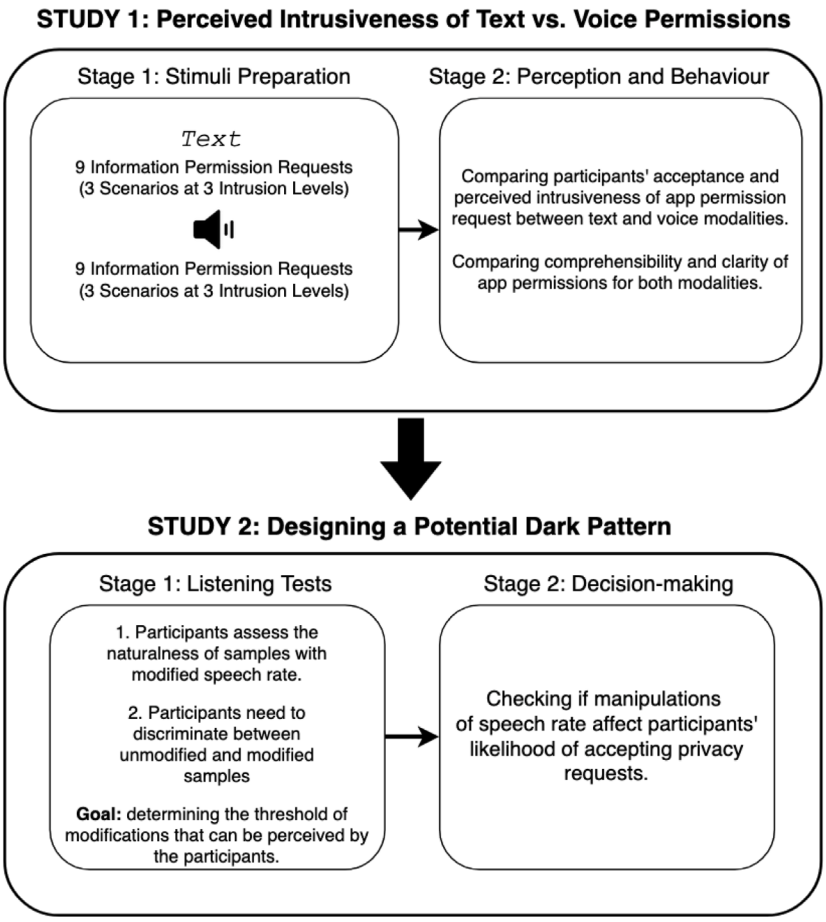


Fig. 1. An overview of experimental structure illustrating the relationship between Study 1 and Study 2.

Table 1
Overview of the systematic construction of permission requests across three different scenarios and intrusiveness level. While the requested information varies by scenario, the purpose of data collection, data sharing, and data retention periods remain the same across scenarios.

Data practice	Low intrusion	Medium intrusion	High intrusion
Information request			
Scenario 1 - Ride Hailer	Pick-up address, First name	Pick-up address, Phone number	Pick-up address, Credit card information
Scenario 2 - Fitness Buddy	Fitness goals, Gender	Fitness goals, Weight	Fitness goals, Medical history
Scenario 3 - Kitchen Wizard	Dietary restrictions, First name	Dietary restrictions, Social media profile	Dietary restrictions, Credit card information
Purpose of data collection	Provide service (e.g., find a ride, provide training recommendation, provide recipe suggestions), Use for current service only	Provide service, Communicate special offers	Provide service, Any other use deemed helpful
Shared With...	Nobody	European third-parties	Third-parties worldwide
Data retention period	Deleted after the interaction	Deleted after 6 months	Deleted after 10 years

is presented in the permission request. The impact of intrusiveness on user acceptance has been explored in the context of mobile application permission requests, where it was found that app value often outweighs the influence of intrusiveness and privacy concerns (Wottrich et al., 2018). To investigate the role of intrusiveness on users' likelihood to accept permission requests in voice-based and text-based scenarios, we varied not only the information requested but also the purpose of data collection, data sharing practices, and data retention periods.

We chose the type of information requested for each level of intrusiveness based on previous research on information sensitivity (Schomakers et al., 2019). Moreover, a pan-European survey found that user preference for Internet Service Providers decreased with longer data retention periods, ranging from one month to five years (Potoglou et al., 2017). Consequently, we increased retention periods with increasing intrusiveness from no retention to retention of up to ten years. Furthermore, previous research indicates that risk perception increases when data is being shared with or sold to third parties and as the geographical scale of sharing expands (Emami-Naeini et al., 2021; Potoglou et al., 2017). Thus, we followed a similar principle and increased the geographical scale of data sharing with each level of

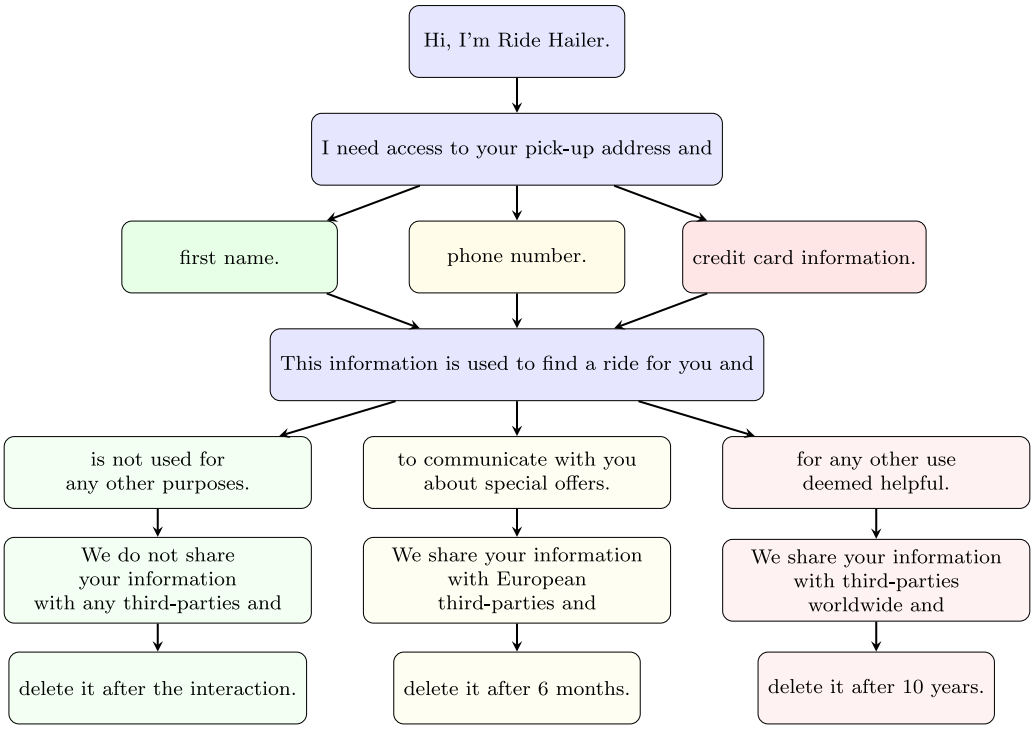


Fig. 2. Permission Requests for Ride Hailer for the three intrusiveness levels — low intrusion in green, medium intrusion in yellow and high intrusion in red.

intrusiveness. Finally, we adjusted the purpose of the data collection for each level of intrusiveness based on previous research. Bhatia and Breaux (2017) distinguish between six categories of data purpose based on an analysis of privacy policies from the shopping domain (i.e., service, legal, communication, protection, merger, and vague purpose). Moreover, studies suggest that users are more likely to accept purposes that benefit themselves or the greater good, rather than service providers or third-party vendors (Kyi et al., 2023; Shih et al., 2015). Finally, a study on privacy preferences in smartphone apps found that providing no purpose increases users' willingness to disclose data, whereas providing a vague purpose raises privacy awareness and reduces disclosure intention (Shih et al., 2015). Following these insights, we moved from service purpose for the least intrusive permission request, to communication purpose for medium intrusive requests and to vague purpose for the most intrusive permission request. Table 1 provides an overview of our systematic construction of permission requests. The final permission requests for the three intrusion levels for the "Ride Hailer" application are shown in Fig. 2.

4.1.2. Audio generation

The generation of speech from text is central to both Study 1 and Study 2. We generate the speech using a Text-to-Speech (TTS) synthesis system that enables controlled manipulation of speech characteristics, such as speech rate, which was essential for implementing voice-based deceptive patterns in Study 2. The same TTS system was used consistently across both studies and is described in detail below.

Text-to-Speech synthesis produces a speech output from a given text input (Taylor, 2009). The goal is to synthesize speech that is both intelligible and sounds natural to human listeners. Due to the application of deep learning and neural networks, recent TTS systems are able to synthesize highly natural speech (Ren et al., 2021; Elias et al., 2021; Huang et al., 2022; Nguyen et al., 2023; Tan et al., 2024; Ju et al., 2024; Eskimez et al., 2024; Shibuya et al., 2024).

The neural-network-based TTS used for our experiments consists of three parts: text analysis, acoustic model, and neural vocoder. During text analysis, the input text is processed and converted into a phoneme transcription, which serves as a representation of the speech sounds corresponding to the text. Phonemes are the smallest unit of sound in language that help to distinguish one word from another in a given language, e.g., *d* and *t* in the English words *bad* (/bæd/) and *bat* (/bæt/). The acoustic model is realized as a neural network which uses phoneme transcriptions to create an intermediate representation of speech features. Mel spectrograms are used as intermediate representations since they capture the intensity of frequencies over time and are scaled to match human perception of sounds. These mel spectrograms are finally converted by the neural vocoder model into a speech waveform.

For our experiments, we synthesized text prompts using an internal TTS system. The architecture of our acoustic model is based on ForwardTacotron (Schäfer et al., 2020), with extensions similar to the model by Zalkow et al. (2023). Our acoustic model is trained using speech recordings (resampled to a sampling frequency of 22,050 Hz) and mel spectrograms (80 bands, hop size of 256 samples, block size of 1024 samples). One challenge in TTS is that the dimension of the input features (textual information) is much lower than the dimension of the intermediate features (mel spectrogram) and the output features (speech waveform). Therefore, an approach for mapping between these feature dimensions is required. Our model belongs to the family of parallel models which utilize a duration prediction module for this purpose. The duration prediction module estimates the number of mel spectrogram frames corresponding to each phoneme to ensure that each sound in the generated speech has the appropriate length, e.g., for the phoneme transcription /bæt/, the duration predictor could estimate that two mel spectrogram frames should be used for the /b/, four frames should correspond to /æ/ (vowels often have a longer duration) and the /t/ receives one mel spectrogram frame. These predicted phoneme durations are fed into a length regulator, which maps features from a low-dimensional text space to a higher-dimensional mel spectrogram space. This is done by repeating the features on the phoneme level in a non-equidistant way based on the predicted phoneme durations.

When synthesizing speech with this model, the predicted phoneme durations can be deliberately modified. By scaling the phoneme durations, we can change the duration of specific phonemes in the synthesized speech. Specifically, by increasing the predicted durations of all phonemes belonging to a word, we increase the number of mel

spectrogram frames for this word, which leads to a higher duration of this word in the synthesized speech. This feature allowed us to control the duration of specific words in our permission requests. As neural vocoder, we use StyleMelGAN (Mustafa et al., 2021).

We trained the acoustic model on a multi-speaker dataset. For the following experiments, we select the voice from the Hi-Fi dataset (Bakhturina et al., 2021) for synthesis. This voice belongs to a female British English speaker. The recordings of this speaker have an average pitch of 210 Hz (SD = 30 Hz), which approximates an average female voice (≈205 Hz) (Holmberg et al., 1988; Simpson, 2009; Puts et al., 2006). The speech rate in the dataset is approximately 4.8 syllables per second (199 words per minute). Examples of these samples can be found in the supplementary material.

4.2. Methodology

4.2.1. Experimental setup

The study implemented a 2 × 3 factorial design, with the first factor including **text** or **voice**-presentation of the permission request, and the second factor including three levels of intrusiveness. As described in Section 4.1.1, the permission requests and intrusiveness levels were informed by previous research. In total, this resulted in six conditions. To mitigate the effect of specific app descriptions on the results, we present three different app descriptions, i.e., "Ride Hailer", "Fitness Buddy" and "Kitchen Wizard", in random order on all levels of modality and intrusiveness. In our analysis, we add the type of app as a control variable. The study was preregistered.[9]

For the study, we implemented the online setup on the LimeSurvey platform,[10] Following the study protocol, participants were invited to read a description of the imaginary voice-based app. Afterwards, we asked participants to read or listen to the app permission request presented by the system in text or voice. They were allowed to reread and replay the audio as often as needed. Each participant interacted with only one of the six conditions. Participants were asked to rate how much they would accept/not accept the request, i.e., *How likely are you to accept this permission request?* followed by questions on intrusiveness, clarity, comprehension and privacy concerns. We also added two open-ended questions to explore if the information provided was sufficient and if any information was missing. The detailed questionnaire and study protocol can be found in the supplementary material.

4.2.2. Measurements

Acceptance of Permission Requests. To measure the likelihood of accepting the permission request, we adapted (Wottrich et al., 2018) permission acceptance scale, which is based on Bernritter et al. (2016) liking intention scale. Acceptance was measured from 0 ("Not at all") to 100 ("Surely accept") with the slider set to the middle position by default.

Intrusiveness. For measuring the intrusiveness, we followed Wottrich et al. (2018) and adapted the version of Nowak and Phelps (1992) privacy questionnaire. It includes four items, measured on a 7-point Likert scale (from "strongly disagree" to "strongly agree").

Clarity and Comprehension. We assessed clarity and comprehension by adapting five items proposed by Masotina and Spagnolli (2022) in the context of text-based privacy notices. Instead of the original 5-point Likert scale version, we applied a 7-point Likert scale (from "strongly disagree" to "strongly agree") to maintain consistency between scales and because of the better psycho-diagnostic properties of the 7-point scale (Finstad, 2010).

Control Variables. Lastly, we included a measure of privacy concerns as a control variable using the validated version of the Internet Users' Information Privacy Concern scale (IUIPC-8) (Groß, 2023), measured on a 7-point Likert scale (from "strongly disagree" to "strongly agree"). We also used the gender and age of the participants as control variables, as previous studies showed that these factors can affect synthetic voice perception (Levitan et al., 2018; Baumann, 2017; Mazanec and Mccall, 1976; Hall, 1978; Hall et al., 2000).

Open-Ended Questions In addition to the quantitative questions, we included two open-ended questions regarding the information provided in the permission request. We asked participants if they found the provided information sufficient and whether they felt that anything was missing from the permission request.

4.2.3. Participant recruitment

We recruited participants from Prolific using the following criteria: (1) they were based in the United Kingdom, (2) spoke English as their first language, and (3) had an approval rate above 99% on Prolific and had been registered on the platform for at least one year. We selected UK-based participants due to their familiarity with the Standard British accent, as we used a female British English speaker to synthesize permission requests (as explained in Section 4.1.2). Additionally, we preregistered the exclusion criteria for participants who completed the study significantly faster than the average. Specifically, participants were excluded if their completion time exceeded two standard deviations below the study's mean completion time. These constraints were introduced to minimize the risk of including the results of individuals who might not have completed the study up to the required standard. As the study implemented the experimental design (comparing between conditions), we did not apply any additional criteria before data collection. However, we settled on the gender balance option, as previous studies in the voice domain showed the effect of the listener's gender on speech perception (Mazanec and Mccall, 1976; Hall, 1978; Hall et al., 2000). We used the G*Power software tool,[11] to calculate our study's power. To analyze the study's main hypothesis (power of 0.8, medium effect size), we conducted an ANOVA incorporating interaction effects which yielded a required sample size of 269 participants. To ensure that we would have sufficient statistical power even in the event of possible outlier removal, we decided to collect responses from 360 participants (60 per condition).

4.2.4. Ethical considerations

The design and data collection were approved by the Ethical Review Panel of University of Luxembourg. We followed the Prolific Guidelines in compensation rate,[12] therefore, participants were paid at the rate of £9 per hour and the consent form and contact information of the researchers were provided.

4.3. Results

4.3.1. Data cleaning and descriptive statistics

During the procedure, we collected the results from 360 participants. After checking the results' integrity, we calculated the median completion time (211 s) and deducted two standard deviations(113 s). All the results below 113 s were deleted, resulting in 343 answers retained for the main analysis. The sample had the following characteristics: the mean age of participants was 41.53 years (SD 12.4), 172 participants identified as female, 168 as male, and 3 participants did not provide the information. The mean for the privacy concerns scale was 47.73 (SD 5.99, min = 30, max = 56), which can be considered as a rather privacy-concerned sample[13] Preliminary analysis of the data distribution showed a significant violation of the normality assumption (Shapiro–Wilks test on all the scales was significant with $p < .001$) for all the tested scales, therefore, we used a non-parametric approach for our analysis.

[9] https://aspredicted.org/6SK_ZW1.

[10] https://www.limesurvey.org/.

[11] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.

[12] https://researcher-help.prolific.com/en/article/2273bd.

[13] Comparing to Abrokwa et al. (2021), where the population with a similar mean (47.7) were described as being on the "concerned side".

Table 2
Quade nonparametric analysis of variance (ANCOVA) with post-hoc pairwise comparison of groups in request acceptance and perceived intrusiveness.

Acceptance: **Voice-based Request**					Perceived intrusiveness: **Voice-based Request**				
F = 61.55	DFH = 2	DFE = 175	p < .001	Bonf-corr p < .001	F = 58.075	DFH = 2	DFE = 175	p < .001	Bonf-corr p < .001
Pairwise comparison of groups									
Comparison	t	DF	p	Bonf-corr.p	Comparison	t	DF	p	Bonf-corr.p
Low vs. Medium	6.579	175	<.001	<.001	Low vs. Medium	−5.810	175	<.001	<.001
Low vs. High	11.046	175	<.001	<.001	Low vs. High	−10.773	175	<.001	<.001
Medium vs. High	4.571	175	<.001	<.001	Medium vs. High	−5.072	175	<.001	<.001
Acceptance: **Text-based request**					Perceived intrusiveness: **Text-based Request**				
F = 71.863	DFH = 2	DFE = 159	p < .001	Bonf-corr p < .001	F = 113.098	DFH = 2	DFE = 159	p < .001	Bonf-corr p < .001
Pairwise comparison of groups									
Comparison	t	DF	p	Bonf-corr.p	Comparison	t	DF	p	Bonf-corr.p
Low vs. Medium	7.918	159	<.001	<.001	Low vs. Medium	−9.374	159	<.001	<.001
Low vs. High	11.612	159	<.001	<.001	Low vs. High	−14.731	159	<.001	<.001
Medium vs. High	3.581	159	<.001	<.001	Medium vs. High	−5.183	159	<.001	<.001

4.3.2. Acceptance and perceived intrusiveness across intrusiveness levels

We implemented two of Quade's non-parametric ANCOVA models for text and voice-based modalities of permission requests, using age, gender, type of app, and level of privacy concerns as covariates. The results (see Table 2 for details) showed significant differences between the three different requests ranked by the level of their intrusiveness, in both voice and text modality by parameters of acceptance and perceived intrusiveness of the requests, even after family-wise Bonferroni correction. The pair-wise post hoc analysis revealed significant differences between each level of intrusiveness in both modalities. The low level of intrusiveness was perceived as both more acceptable and less intrusive compared to the middle level, and the middle level compared to the high level of intrusiveness. The results support H1 and H2.

4.3.3. Acceptance and perceived intrusiveness across modalities and intrusiveness levels

Following our preregistered analysis plan, we made pair-wise comparisons between acceptance and perceived intrusiveness at each level of intrusiveness of the requests using three U-Mann–Whitney tests (family-wise, Bonferroni corrected for three comparisons). The results (see Table 3) showed significant differences in acceptance between the modalities at the medium levels of privacy intrusiveness.

In such cases, the voice-based request was perceived as less acceptable.[14] However, we did not find significant differences between the modalities at different levels of perceived intrusiveness. The results partly confirmed H3b, but provided no evidence to support H4a or H4b.

To understand the potential impact of different application scenarios on the variables Likelihood to Accept and Perceived Intrusiveness and to assess the robustness of the main analysis, we conducted an additional analysis that included these variables as variables of interest.

Kruskal–Wallis tests revealed no significant differences between scenarios in terms of Likelihood to Accept ($H = 0.776$, $df = 2$, $p = .678$) or Perceived Intrusiveness ($H = 0.482$, $df = 2$, $p = .786$).

Additionally, we performed linear regression analyses including Application and Level of Intrusiveness of Request as predictors in both models. For predicting Likelihood to Accept, the overall model was significant ($F(2, 340) = 32.869$, $p < .001$), but Application was not a significant predictor ($B = 2.805$, $p = .244$). Similarly, for Perceived Intrusiveness, the model was also significant overall ($F(2, 340) = 28.568$, $p < .001$), yet Application again did not significantly predict the outcome ($B = –0.128$, $p = .794$). In both models, Level of Intrusiveness of Request emerged as a significant predictor ($p < .001$).

These results suggest that the specific content of the scenarios likely did not have a significant effect on the study outcomes.

[14] Pre-corrected data showed that we have the same tendency in the low-level intrusiveness case, however, the result did not stand against the correction procedure.

4.3.4. Clarity and comprehension across modalities and intrusiveness levels

We ran three U-Mann–Whitney tests (family-wise, Bonferroni corrected for three comparisons) to measure the differences in clarity and comprehension between modalities at each level of intrusiveness. The results showed no significant differences at any level for both parameters, except for clarity at the higher level of intrusiveness. At this level, the text-based request was perceived as significantly clearer than the voice-based request ($U = 1.923$, $p = .03$). However, this result does not hold up to correction for multiple comparisons. Therefore, we conclude that we did not find evidence to support H5.

4.3.5. Qualitative analysis and results

To analyze the responses to the open-ended questions, we employed affinity mapping (Beyer and Holtzblatt, 1997; Kawakita, 1982). Thereby, we extracted participants' responses to the two questions and within each condition, e.g., text-based and low intrusiveness, together with the application that they interacted with. Adding information about the specific application allowed us to identify potential differences in responses due to the characteristics of the application's permission request. The first author clustered responses within each condition to identify common themes and patterns, while the second author reviewed and confirmed the grouping. The emerging themes, including "Understandability", "Third-Party Requests", "Opt-Out and User Control", "Text-based Permission Preference", were discussed by the two authors and compared among conditions to ensure that unique insights were captured.

The first open-ended question asked users to assess whether the provided information was sufficient (*"Do you think that the information provided in the permission request is sufficient? Please explain your answer"*). The qualitative responses confirm the trend observed in the quantitative results. In both voice and text-based conditions, the number of participants who found the information to be sufficient decreased as the level of intrusiveness increased. In text-based conditions, participants who perceived the permission request as sufficient decreased from 75% in low intrusiveness levels to 44% in medium intrusiveness levels to 30% in high intrusiveness levels. In voice-based conditions, participants who perceived the permission request as sufficient decreased from 80% in low intrusiveness levels to 40% in medium intrusiveness levels to 28% in high intrusiveness levels. At medium and high levels of intrusiveness, participants' main concerns were related to third-party sharing and credit card information. For instance, P42 noted *"it should tell exactly what companies the data will be shared with and include a link where you can find more information on those companies and their reputation"*. In the high intrusiveness condition, five participants noted that the information was perceived as sufficient yet too intrusive. For instance, P3 noted *"I feel that the information provided in the permission request was sufficient for me personally to not wish to use the app"*. Moreover, two outliers indicated trust issues related to

Table 3
Results of U Mann–Whitney test for Acceptance, Perceived intrusiveness, Comprehension and clarity between modalities across different levels of intrusiveness.

Comparison (Text vs. Voice)	N	Mann-Whitney U	p	Bonf-corr p (3 comp.)	Comparison (Text vs. Voice)	N	Mann-Whitney U	P	Bonf-corr p (3 comp.)
Acceptance					Comprehension				
Low	118	1363	**.039**	.117	Low	118	1655	.638	1
Middle	113	1161	**.014**	**.042**	Middle	113	1680	.586	1
High	112	1423	.424	1	High	112	1258	.079	.237
Perceived intrusiveness					Clarity				
Low	118	1799	.33	1	Low	118	1406	.068	.204
Middle	113	1907	.063	.189	Middle	113	1836	.147	.441
High	112	1771	.205	.615	High	112	1923	**.03**	.09

permission requests, with P10 stating, *"It depends on if the information covered everything it needed to if so yes it's sufficient otherwise it's not [...]"*. In both voice and text-based conditions, people provided varied reasons for why they found the information to be sufficient. The two key factors encompassed clarity and understandability and the explanation of data practices such as data collection purpose or data retention period. For instance, in the text-based condition and low intrusiveness level, 16% mention understandability and clear language while 25% found the information provided sufficient due to the explanation of data practices. In the corresponding voice-based condition, 33% discuss data principles in their reasoning and only 9% clarity and understandability aspects. Overall, we identified a tendency for clarity and understandability to be more frequently mentioned in the text-based conditions as compared to the voice-based conditions. For instance, P20 noted for a text-based permission request of medium intrusiveness *"Yes, it was explained clearly and I understood how my data was to be used"*, whereas P17 highlighted for a voice-based permission request of medium intrusiveness *"Yes, it was long enough that you got told what it was for and when it would be deleted but it didn't drone on so people would get bored and possibly misunderstand. Also, the language was plain no 'big words'"*.

When analyzing responses to the second open-ended question (*"What specific information would you like to find in the permission request that is currently missing?"*), we found that in both voice- and text-based conditions participants frequently noted missing details regarding data usage, data retention, and data storage. They expected detailed information regarding data storage, security guarantees, opt-out mechanisms and user controls. For instance, P55 asked for *"information about security and protection"* and P47 noted, *"I'd like to understand how to opt out of data sharing"*. This highlights the need for future studies regarding the information that needs to be included in short and specific consent and permission requests more broadly, as these concerns were raised for both voice- and text-based conditions. Moreover, participants frequently asked for reassurance that data would not be sold to third parties and requested further clarification on data deletion, even at low intrusiveness levels, despite clear data deletion statements. For instance, P84 noted in voice-based and low intrusiveness condition *"It would be helpful to have more information on the amount of time they would withhold the data"* while P5 remarked in text-based and low intrusiveness condition *"they could tell you how long they hold your data for"*. These responses raise questions regarding comprehension of permission requests in general, as we identified these misunderstandings across modalities.

In response to both questions, participants in voice-based conditions with medium or high intrusiveness levels preferred written text or an additional link for more information. One participant (P23) expressed discomfort with AI-generated voices by stating *"using an AI voice did not fill me with confidence at all"*.. P40 expressed concern about how one could learn more by reading a privacy policy when the information is provided through an audio recording. This suggests that while voice-based permission requests can convey straightforward information, they may be less suitable for more intrusive requests that might require additional clarification. Finally, differences between applications were minimal with respect to the sufficiency of information provided or missing information. Yet, in the high intrusiveness condition, participants raised more questions about their credit card information than about medical information, e.g., "The permission request made me feel uncomfortable by asking for my credit card details — it did not explain what it is do with these which did not make me feel secure". (P38) This finding is similar to the recent (Desai et al., 2024) study, where some participants were cautious regarding providing financial information to a voice interface and found it unnerving and unwise. However, it should be noted that different people or event the same person may have different sensitivity levels towards the same subject depending on their current personal circumstances such as their health or financial situation (Schuetzler et al., 2018).

4.4. Discussion

The results showed that people can clearly distinguish between different levels of app intrusiveness in both modalities, aligned with decreased acceptance rates for higher levels of intrusiveness. This suggests that short and understandable permission requests can be translated into the voice domain. Nevertheless, through our qualitative analysis, we identified misunderstandings related to data deletion and third-party sharing in both modalities, raising questions about comprehension of permission requests in general. Thereby, qualitative analysis can be beneficial in capturing nuances that are not reflected by quantitative scales.

We did not find evidence that people perceive intrusiveness differently for text or voice-based requests. However, there is evidence that people tend to accept voice-based permission requests less than text-based requests, at least at the medium level of app intrusiveness. We hypothesize that at the highest level of intrusiveness, the content of the request may become so dominant that it overrides the influence of how information is presented. However, further research is needed to determine whether differences in modality emerge in larger-scale evaluation. Generally, our finding raises questions regarding users' more critical attitude towards voice. This might be due to the general novelty of voice-based permission requests, as well as a perceived lower level of control of voice communication.

In line with this, our qualitative analysis highlights that users prefer written permission requests in medium and high intrusiveness levels and that synthetically generated voice can make them feel uncomfortable.

5. Study 2: Effects of voice deceptive patterns on user perception and decision-making

The results from Study 1 (presented in Section 4.3) indicate that voice-based permission requests are perceived differently compared to screen-based requests. As discussed in Section 2.2, akin to deceptive patterns in the visual domain, voice deceptive patterns could nudge users into disclosing personal information through manipulations of prosodic attributes, such as speech rate or pitch range (Owens et al.,

2022). To explore this further, we designed experiments to investigate how variations in speech rate affect users' decisions to accept or decline permission requests. Specifically, we synthesized the phrase, "Please choose Decline or Accept" while manipulating the speech rate of either the word "Decline" or "Accept". As previous studies showed the potential of speech rate to channel priming effect response (Jungers and Hupp, 2009; Tooley et al., 2018), we decided to test the possibility of affecting user choice by this manipulation by making one of the options more audibly salient (slower) to the participants. Therefore, we expected an increase in acceptance rates in the condition of the shortened "Decline" option and a decrease in acceptance rates in the condition of the shortened "Accept" option. Response order effects can influence decisions differently in text and voice formats (Krosnick and Alwin, 1987). Primacy effects, i.e., earlier options are chosen more often, are typically observed in written surveys, while recency effects, i.e., options are selected more often when presented later, are more common in oral formats (Krosnick and Alwin, 1987). However, such recency effects in orally presented dichotomous questions appear relatively infrequently, occurring in only 19.2% of cases (Holbrook et al., 2007). Given this and our main focus on investigating voice-deceptive patterns, we used a fixed order ("Please choose Decline or Accept") for Study 2.

In Study 2, we seek to answer the following research questions:

Research Question 4 (RQ4): Do changes in the speech rate of "Accept" or "Decline" affect the acceptance or declining rate of the app permission request? More specifically, we hypothesized that increasing the speech rate of either "Accept" or "Decline" would make the opposite option more salient, prime participants to choose it, and, in turn affect the acceptance rate. Consequently, a faster speech rate for "Accept" would lead to more "Decline" choices and vice versa **(H6)**.

Research Question 5 (RQ5): Do changes in speech rate of "Accept" or "Decline" affect the emotional response in the post hoc users' evaluation of the request; we hypothesized that if users will perceive the manipulation attempts in the phrases, they will have more negative feelings to the requests **(H7)**.

Research Question 6 (RQ6): Will users recognize the difference in speech rate of "Accept" or "Decline" as an important factor, affecting their decision-making? We hypothesized that in this case, the level of importance of the assistant's speech rate would be significantly lower in the control condition compared to the manipulated conditions **(H8)**.

5.1. Pre-study: Designing potential voice deceptive patterns

Before investigating the effect of our deceptive patterns on users' perceptions and decision-making, we conducted a series of listening tests to identify suitable changes in speech rate that preserve the naturalness of the signal while remaining noticeable to users when listening to the isolated signals. The purpose of this pre-study was to select appropriate speech rate manipulations for Study 2. Our aim was to ensure that the manipulated versions sounded natural while remaining perceptible when listened to in isolation. By first evaluating user perception of these manipulations in a controlled setting, we were able to ground the design choices of Study 2 in experimentally validated parameters and provide insights into how potential deceptive patterns can be constructed and detected in voice user interfaces.

In this pre-study, we seek to answer the following research questions:

Research Question 7 (RQ7): How do users perceive the naturalness of audio samples that have been prosodically manipulated? More specifically, we hypothesized that perceived naturalness of audio samples will decrease as the level of prosodic change increases compared to the reference signal **(H9)**.

Research Question 8 (RQ8): How well can users discriminate audio samples that have been prosodically manipulated from a non-prosodically manipulated reference signal? We hypothesized that accuracy in discriminating audio samples will increase as the level of prosodic change increases **(H10)**.

5.1.1. Experimental setup

Based on the decision-making sentence "Please choose 'Decline' or 'Accept'", we manipulated the speech rate of the words "Accept" or "Decline". We selected a range of speech rates with step rates between 5% and 20% around a standard rate of 100%, i.e., 80%, 85%, 110%, 120%. In addition, we included one signal with an extreme decrease, i.e., 60% speech rate, to serve as a lower anchor for our listening test on naturalness. Lower anchors are common practice in listening tests to identify outliers in participants' ratings (International Telecommunication Union, 2015).

First, we aimed to investigate the naturalness of prosodically manipulated speech signals. We asked participants to rate the naturalness of speech signals where speech rates of the words "Accept" or "Decline" had been manipulated. All stimuli with the same modified word were shown on the same test page together alongside a non-manipulated reference signal. Naturalness was rated on a scale from 0 to 100 with distinct labels to ensure consistent use of the scale, i.e., Bad (0–19), Poor (20–39), Fair (40–59), Good (60–79), and Excellent (80–100). Participants were allowed to switch between samples during the rating process.

Second, we aimed to determine whether participants could reliably discriminate between prosodically manipulated signals and the reference signal. A discriminatory test design was used in which participants were presented with two audio samples, i.e. one "non-prosodically manipulated" and one "prosodically manipulated" audio and asked to choose which sounded more like a given reference. The reference audio remained the same across trials and always represented the non-manipulated version to reduce the number of tested combinations. They also had the option to choose "I don't know" in cases of uncertainty. This option was included to reduce the likelihood of random guessing and to assess which samples participants could confidently distinguish. Playback of each sample was limited to three times to balance the possibility of missing differences on the first trial while restricting infinite trials.

Participants for both tests were internally recruited from one of the authors' institutions while ensuring that they had a proficient level of English. The tests were released two weeks apart with a break in between to minimize the chances of participants recalling the audio samples and their associated ratings. Both tests were implemented using the webMUSHRA framework (Schoeffler et al., 2018). The test layouts are shown in Fig. 3.

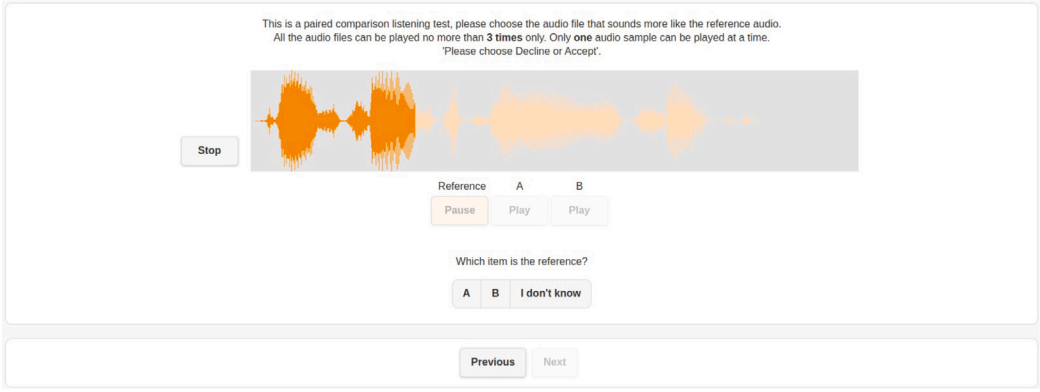
5.1.2. Results

Descriptive statistics. For the first listening test, we collected data from 20 participants balanced by gender (10 female and 10 male) with a mean age of 28.30 years (95% CI 26.32, 30.28). We did not find any outliers based on two standard deviations for completion time and kept all 20 results for the analysis. A Shapiro–Wilk test ($p < .001$) indicated a significant deviation from normality, such that non-parametric tests were used for further analysis.

For the second listening test, we collected data from 26 participants (8 female, 17 male, and 1 other) with a mean age of 33.50 years (95% CI 29.99, 37.01). Again, we did not find any outliers based on two standard deviations for completion time and kept all results.

Differences in naturalness. Fig. 4 shows results for the listening test on naturalness of our modified samples. We conducted Friedman tests to identify differences in naturalness across speech rate manipulations (as shown in Table 4). The results showed significant differences between the reference condition and the speech rate manipulations applied to "Accept" and "Decline", respectively. We further conducted Wilcoxon Signed-Rank tests for pairwise post-hoc analysis between samples with prosodic manipulation and the reference signal. The test revealed significant differences between our chosen lower anchor and the reference signal even after family-wise correction ($W = 6.0$, corr-p = .0027 for "Accept" and $W = 7.5$, corr-p = .0034 for "Decline"). Naturalness

(a) Web interface for the listening test on naturalness.



(b) Web interface of the discriminatory listening test.

Fig. 3. Web interfaces of the two listening tests: naturalness (a), and discriminatory (b).

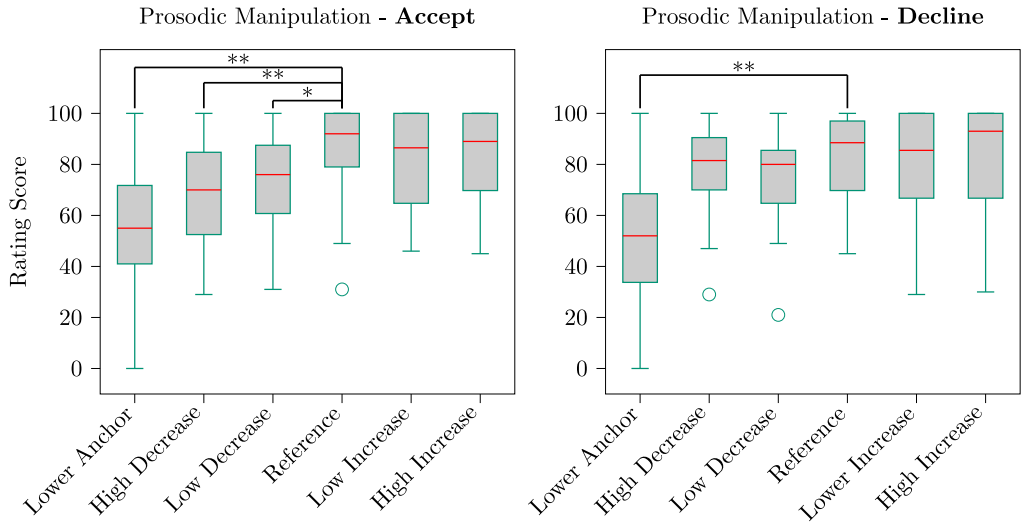


Fig. 4. Rating Scores of users in the naturalness test for speech rate changes of "Accept" and "Decline". Horizontal lines indicate statistically significant differences between conditions and the reference signal as determined by pairwise comparison with $p < .05$ (*) and $p < .01$ (**). A score of 100 indicates "excellent" naturalness, while 0 indicates "bad".

ratings for high and low decreases in speech rate also significantly differed from the reference for "Accept" ($W = 16.5$, corr-p = .0078 for high decrease and $W = 24.0$, corr-p = .0370 for low decrease) but not for "Decline".

Differences in discriminatory accuracy. For each condition, we computed discriminatory accuracy depending on whether participants' ratings were correct, incorrect or undecided. Fig. 5 provides an overview of the rating distribution across conditions. We found that only the

Table 4
Friedman test and pairwise comparison of naturalness ratings between conditions and reference.

Prosodic manipulation: **Accept**				Prosodic manipulation: **Decline**			
Q = 34.21	DF = 5	p < .0001	Bonf-corr p < .0001	Q = 30.86	DF = 5	p < .0001	Bonf-corr p < .0001
Pairwise comparison of groups							
Comparison	W	p	Bonf-corr.p	Comparison	W	p	Bonf-corr.p
Lower anchor vs. Reference	6.0	.0005	.0027	Lower anchor vs. Reference	7.5	.0007	.0034
High decrease vs. Reference	16.5	.0016	.0078	High decrease vs. Reference	73.0	.5859	1.0
Low decrease vs. Reference	24.0	.0074	.0370	Low decrease vs. Reference	46.5	.0892	.4460
Low increase vs. Reference	49.5	.3383	1.0	Low increase vs. Reference	66.5	.9381	1.0
High increase vs. Reference	50.0	.5699	1.0	High increase vs. Reference	60.0	1.0	1.0

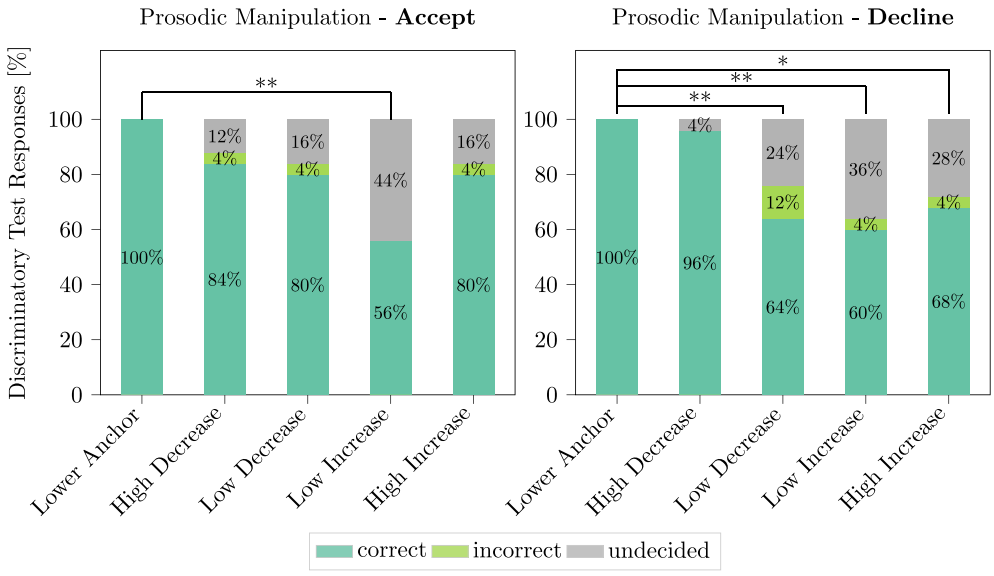


Fig. 5. Percentage of participants who correctly and incorrectly identified the reference speech signal or chose "I don't know" for speech rate changes of "Accept" and "Decline". Horizontal lines indicate statistically significant differences between conditions and the reference signal as determined by pairwise comparison with p < .05 (*) and p < .01 (**).

Table 5
Fisher's exact test for pairwise comparison of discriminatory accuracy between conditions and lower anchor. As the anchor group showed no incorrect responses, the odds ratios result in 0.0 for all comparisons.

Prosodic manipulation: **Accept**				Prosodic manipulation: **Decline**			
χ^2 = 19.77	DF = 8	p = .01	Bonf-corr p = .02	χ^2 = 23.68	DF = 8	p = .003	Bonf-corr p = .005
Pairwise comparison of groups							
Comparison	Odds ratio	p	Bonf-corr.p	Comparison	Odds ratio	p	Bonf-corr.p
High decrease vs. Lower anchor	0.0	.11	.44	High decrease vs. Lower anchor	0.0	1.0	1.0
Low decrease vs. Lower anchor	0.0	.05	.20	Low decrease vs. Lower anchor	0.0	.002	.006
Low increase vs. Lower anchor	0.0	.0002	.001	Low increase vs. Lower anchor	0.0	.0006	.003
High increase vs. Lower anchor	0.0	.05	.20	High Increase vs. Lower Anchor	0.0	0.004	0.02

lower anchor was correctly discriminated from the non-manipulated audio by all participants, while most people were undecided in both manipulated conditions when there was only a slight increase in speech rate. Further, Chi-Square tests were performed to identify differences between conditions. As significant differences were found for both prosodic manipulations, i.e., "Accept" and "Decline"., we conducted Fisher's Exact tests to explore pairwise differences. Detailed results are shown in Table 5. As we aimed to determine speech rate variations that can be confidently recognized by humans, we used the lower anchor as a reference distribution for pairwise testing. This choice was motivated by the fact that the lower anchor was consistently and correctly recognized by all participants. We then compared the distribution of every other condition to that of the lower anchor. Due to the usage of Fisher's Exact test, we combined the incorrect and undecided ratings to construct a 2 × 2 matrix. The results suggest that only a slight increase in speech rate for "Accept" showed significant differences from the lower anchor distribution after the Bonferroni correction. In contrast, slight decrease, slight increase, and high increase showed significant differences compared to the low anchor for "Decline".

5.1.3. Discussion

Overall, the analysis showed that all manipulations except the chosen lower anchor were rated as at least "Good" with respect to naturalness with median values above 84 (SD = 20). Nevertheless, significant differences were found between the reference signal and decreased speech rates of "Accept". This suggests that decreases in

speech rate are problematic for preserving the naturalness of the speech signal. In contrast, increases in speech rate did not lead to significantly different ratings from the reference signal. Moreover, significant differences for decreases in speech rate of "Accept" but not "Decline" suggest that additional factors such as word placement, word structure and inherent capabilities of the TTS system, such as default speech rate, might also impact the perception of naturalness. Our investigations on noticeable differences in speech rate changes unsurprisingly showed that bigger changes are easier to discriminate than subtle ones. Yet, the discriminatory test suggests that changes to "Decline" were less likely to be correctly identified. These findings align with the naturalness test results, where changes to "Decline" were considered less problematic than changes to "Accept". Future research could investigate which additional factors might impact naturalness and noticeable differences, and what role they play in the construction of potential voice deceptive patterns.

The results of the pre-study were further used to inform the design of Study 2, ensuring that changes in speech rate are perceived as both natural and noticeably different from the reference signal. The naturalness test suggested that both low and high increases in speech rate were suitable candidates. However, the discriminatory test showed higher error rates or undecided responses for the low increase condition for both "Accept" (44% undecided) and "Decline" (40% undecided or incorrect). Therefore, we chose the high increase in speech rate (20% faster) for Study 2.

5.2. Main study: Methodology

As discussed in the previous section and based on the pre-study, we chose the high increase in speech rate to evaluate its impact on people's decision-making and perception. To evaluate the effect of voice deceptive patterns on user perception and decision-making, we combined the permission requests discussed in Section 4.1.1 with the speech rate manipulated sentence on accepting or declining. We will describe our experimental setup in more detail in the following section.

5.2.1. Experimental setup

The study implemented a 3 × 3 factorial design, with the first factor including the permission request's acceptance sentence with **increased "Accept"**, **increased "Decline"** or **control**, i.e., original speech rate for both words, and the second factor including the three levels of intrusiveness. We selected only one application for this study, namely "Fitness Buddy", and its corresponding permission requests (as described in Section 4.1.1 and shown in Table 1). This decision was based on the qualitative results, which indicated that differences between applications were minimal, yet the fewest concerns were raised in the second scenario for the highest intrusiveness level. Thus, we identified the "Fitness Buddy" application as most appropriate for evaluating effects across all intrusiveness levels. The study was preregistered.[15]

Similarly to Study 1, we implemented the setup on the LimeSurvey platform. Following the study protocol, participants were invited to read a brief description of the imaginary voice application. After that, we asked participants to listen to the app permission request presented by the system and choose either "Accept" or "Decline" according to their preferences by pushing the corresponding button. To avoid learning and transfer across conditions, each participant interacted with only one of the nine conditions. Following that, we asked questions about participants' regrets and satisfaction with their decisions, as well as questions about their privacy concerns, their decision-making style, and general demographic information. We also asked them about the importance of speech rate in the "Accept" and "Decline" words to their decision.

[15] https://aspredicted.org/N2X_44Q.

5.2.2. Measurements

Acceptance of privacy permissions. To measure the acceptance of permissions, we used binary options "Accept" and "Decline" presented in random order to control for order effect.

Too Much Choice Scale (TMC). To measure possible regrets as well as satisfaction after executing the choice, we used the TMC scale based on Korff and Böhme (2014). The scale measures the post-choice emotional reaction in the context of privacy settings (7 questions, measured on a 7-point Likert Scale with anchors on both sides).

Privacy Concerns. To control for participants' privacy concerns, we again used the validated Internet Users' Information Privacy Concern scale (IUIPC-8), described in Section 4.2.2.

Rational and Intuitive Decision-Making Styles Scales. Further, we controlled for the intensity and preferences of decision-making styles by applying the Rational and Intuitive Decision Styles Scale, developed by Hamilton et al. (2016). The scales measure the extent to which participants have rational or intuitive approaches to making decisions. The scale consists of two subscales, i.e., rational and intuitive decision-making style, with each of them including five questions, measured on a 5-point Likert scale (from "Strongly Disagree" to "Strongly Agree").

Technological Savviness scale. Finally, we used the overall technological savviness scale proposed by Renz et al. (2023) to control for perceived savviness of technologies. The scale consists of one question, i.e., *"I like testing the functions of new technical systems"*, measured on a 6-point Likert scale (from "Strongly Disagree" to "Strongly Agree").

Perceived importance of speech rate manipulation We used a set of 7-point Likert scales (from 1 - "not at all contributed to the decision" to 7 "extremely contributed to the decision") to assess the perceived importance of different factors for making the decision about accepting or declining permission. In the presented study, we only analysed the results of the question about perceived importance of speech rate in "Accept" and "Decline".

In addition to this, we collected data about the gender, age, and educational level of the participants.

5.2.3. Participant recruitment

We ran Study 2 on Prolific, using the same selection criteria as in Study 1 (see Section 4.2.3). We collected 629 full answers.

5.2.4. Ethical considerations

The design and data collection were approved by the Ethical Review Panel of University of Luxembourg. We applied the fair compensation principle of the Prolific studies (participants were paid at the rate of £9 per hour) and provided the consent form and contact information of the researchers. To ensure transparency, we created a debriefing page in our questionnaire specifying the main purpose of the study (effects of speech rate on decision-making).

5.3. Main study: Results

5.3.1. Data cleaning and descriptive statistics

We preregistered exclusion criteria, which included failing one of the two attention checks implemented in the questionnaire: the questions asked to choose a specific option, e.g. "somewhat disagree" as the answer to the question. However, when we began the data analysis, we also checked the time spent on the page where participants were required to evaluate a voice sample. Since each audio sample was nearly 18 s long, we excluded submissions where the time spent on this page was less than 18 s. This exclusion policy was based on the assumption that participants did not listen to the full sample and, therefore, could not provide reliable data. The final sample after the data cleaning consisted of 594 submissions.

The sample had the following characteristics: the mean age of participants was 44.46 years (SD = 13.52), 296 participants identified as female, 295 as male, and 3 participants did not provide the information. The mean for the privacy concerns scale was 47.60 (SD

Table 6
Permission acceptance rate across different levels of intrusiveness and speech rate manipulations.

Condition	Acceptance rate
Control condition	
Low intrusiveness	87,30%
Medium intrusiveness	39,70%
High intrusiveness	25%
Faster speech rate in "Decline"	
Low intrusiveness	89,20%
Medium intrusiveness	44,30%
High intrusiveness	31,80%
Faster speech rate in "Accept"	
Low intrusiveness	93,80%
Medium intrusiveness	55,80%
High intrusiveness	23,30%

Table 7
The χ^2- evaluations of differences between "Accept" and "Decline" distributions across the conditions of severity.

Intrusion level	N	df	χ^2	P	Bonferroni-corrected p
Low	224	2	1.936	.380	1
Medium	191	2	3.291	.193	.579
High	179	2	1.041	.594	1

= 6.17), the median level of technological savviness was 4 out of 6, and the median level of education was a Bachelor's degree. Since the preliminary analysis of the TMC scale showed a significant violation of the normality assumption (Shapiro–Wilks test was significant with p < .001), we decided to proceed with a non-parametric analysis model. The permission acceptance rate is presented in Table 6.

5.3.2. Effect of speech rate manipulation on acceptance rate

We ran three separate CHI-square analyses on each level of intrusiveness of app permissions, Bonferroni corrected by three comparisons. We did not find significant differences between conditions on any level of intrusiveness. The results are presented in Table 7.

To measure the presence of the effect on the level of the full dataset, we proceed to a binary logistic regression model using the parameters of intrusiveness levels and manipulation levels as variables. We run two regression models, one using the theoretically expected order of levels (from shortened "Accept" to shortened "Decline" with the control category in the middle) and one using the the group with most observations (Control group) as a reference category.

Comparing the models, we found that the model with the control category as a reference point performs slightly better than the theoretically expected model (Nagelkerke R Square = .356 vs. .350). Taken together, it was suggested that both manipulation levels lead to an accept decision significantly more often than the control level. Therefore, we decided to follow the approach used in previous research (Dubiel et al., 2024b) and test the general hypothesis of whether there are differences in acceptance between manipulated and non-manipulated levels. To do so, we merged the manipulated categories into one and ran the regression model using the new binary category and level of intrusiveness. The model provided significant results for both predicting variables of manipulativeness and intrusion levels (χ^2 = 182.663, p < .001; Nagelkerke R Square = .355, Intrusiveness β = −1.560, p < .001, Control vs. Manipulation β = .401, p = .047).

Following the preregistration, we then investigated whether any other control variable (age, gender, rational and intuitive decision-making style score, technological savviness) affected the number of accepts or declines. To check for multicollinearity, we run Spearman Rho correlation analysis. We found that the privacy concerns scale is significantly correlated with five out of the seven other predictors, so we decided to exclude it from the regression model. We also found a significant correlation between rational and intuitive decision-making style (rho = –.247), so we decided to exclude intuitive decision-making style from the model as well. Following the stepwise approach to remove non-significant predictors from the models, we ended up with the model presented in Table 8. The model explains 42% of the variance and has a non-significant result in the Hosmer–Lemeshow test, indicating a considerable goodness-of-fit. The final model suggested that technological savviness, rational decision-making, and experimental conditions, i.e., levels of intrusiveness and manipulated vs. non-manipulated speech rate conditions, are important predictors of acceptance of permissions. Moreover, while the more rational decision-making style decreases the chance of accepting permissions, technological savviness increases it. The correlation table and alternative regression models are presented in the supplementary material.

In summary, we did not find significant evidence supporting the original H6 hypothesis, i.e., increasing the speech rate of either "Accept" or "Decline" makes the opposite option more salient, acts as a prime and consequently affects acceptance rate. However, we found some evidence that adding speech rate imbalance within the privacy permission clause can sway the choice of accepting and declining permission towards acceptance.

5.3.3. Effect of speech rate manipulation on choice regrets

The mean of results of TMC scale showed M = 18.05 (SD = 5.136) (scale maximum value of dissatisfaction is 42) which demonstrate rather satisfied opinion about possibility to make a privacy choice across conditions.

We ran a Quade non-parametric ANCOVA model to determine the differences between the control and manipulated groups for emotional responses on the TMC scale and using the intrusiveness of the condition as a covariate. We did not find significant differences between any of the conditions (F = 0.042, DFH = 2, DFE = 591, p = .959). Therefore, we did not find evidence supporting H7 that manipulation attempts affect participants' emotional response.

5.3.4. Differences in perceived importance of speech rate manipulation

The mean of the answers to the question about perceived importance of the speech rate in "Accept" and "Decline" was 3.16 (SD = 1.815), which indicated a rather moderate level of perceived contribution of this factor to decision making (less than middle point of the scale). It can indicate that users are taking this presentational parameter into account, but did not perceive it as a main factor for their decision-making.

To determine if there are any differences perceived importance of speech rate in "Accept" and "Decline" clause, we ran Quade non-parametric ANCOVA with intrusion level as a covariate. We did not find significant differences between any of the conditions (F = 1.427, DFH = 2, DFE = 591, p = .241). Therefore, we did not find evidence supporting H8 that people in manipulated condition perceived the manipulated speech rate in "Accept" and "Decline" clauses as affecting their decision.

5.4. Main study: Discussion

The results of Study 2 demonstrated that while we did not find evidence for a direct link between speech rate manipulation and change in accepting or declining app permissions, there is evidence that both of them can shift the decision pool towards acceptance. A possible explanation could be that in the setting where "Accept" and "Decline" are present in equal speech rates, more reflection is triggered as both options have relatively the same prominence. However, in case of shortening the duration of one of the words, the presentation of the choices becomes less balanced, so users rely on their existing heuristics in accepting or not accepting a certain set of permissions. The habituation component can also explain the significant positive

Table 8
Binary regression model (combined manipulation group, minimized number of significant predictors).

Model summary						
Omnibus test of model coefficients: p < .001						
Nagelkerke R square = .421						
Hosmer and Lemeshow test: chi-square = 13.837 df = 8 p = .086[a]						
Variables in the equation						
	B	S.E.	Wald	df	Sig.	Exp(B) size-effect
Level of intrusiveness	−1.614	0.140	132.078	1	<.001	0.199
Technological savviness	0.428	0.086	25.038	1	<.001	1.535
Rational decision-making	−0.159	0.038	17.459	1	<.001	0.853
Manipulation groups (Combined manipulation vs. Control)	−0.446	0.213	4.374	1	.036	0.640
Constant	5.522	0.929	35.320	1	<.001	250.179

[a] Non-significant test represent goodness-of-fit.

role of technological savviness in accepting permission. As experience with different technological systems and applications grows, the act of accepting permissions becomes more automatic and less reflective. However, as the results did not match what was theoretically expected, we suggest interpreting our findings with caution. Similarly to the recent study of Dubiel et al. (2024b), we did not find evidence showing users' awareness of manipulative attempts in voice-based systems. We also did not find evidence that the manipulative attempts affected choice regrets afterwards.

6. General discussion and design implications

Our series of studies makes valuable contributions to the field of voice-based permission requests and deceptive patterns in this domain. First, by comparing text-based and voice-based permission requests both quantitatively and qualitatively, we showed that users were less likely to accept voice-based permission requests compared to text-based permission requests for medium levels of app intrusiveness. Moreover, in their qualitative responses, users expressed scepticism and discomfort with AI-generated voices and voice-based permission requests more generally. Through a pre-study, we attempted to design voice deceptive patterns by varying the speech rate of only one word while maintaining naturalness and ensuring noticeable differences to the original speech signal. During this design process, we highlighted that naturalness and noticeable differences are not only due to changes in speech rate but might also be influenced by word structure, such as syllable structure or word-internal phonetic structure, particularly differences in final consonants, like the plosive /t/ in "accept" versus the nasal /n/ in "decline", and TTS capabilities. This urges the need to investigate variations of voice deceptive patterns, both linguistically as well as prosodically, and to identify generalizable patterns to enable detection. Finally, we analyzed the effect of these patterns on users' perception and decision-making. While we did not observe effects as expected, our results indicate that changes in speech rate can affect users' choices, i.e., users were more likely to accept permission requests in both manipulated conditions than in the control condition.

Voice-based permission requests offer a unique way to convey information to users while avoiding the negative consequences of modality switching. However, this modality also presents several potential pitfalls. To address these, we now discuss design recommendations for clear, concise, and responsible implementation.

6.1. Content of voice-based permission requests

One of the primary challenges with voice-based permission requests is the transient nature of speech. Unlike written text, which can be easily reviewed and referenced, spoken information is fleeting. Current approaches to voice-forward consent will time out after eight seconds asking the user to switch modalities and consent through the mobile application.[16] Our qualitative results in Study 1 (discussed in Section 4.3.5) revealed that users were missing certain information independent of the modality. However, clearly communicating more complex requests in the voice domain, such as those including third-party data sharing, may be difficult to convey in a single verbal message. It may be hard for users to process large amounts of information presented verbally, especially when the content is complex or unfamiliar (Baddeley, 1992; Seymour et al., 2022). Future research is needed to identify what information needs to be presented in voice-based permission requests to ensure both legal compliance and usability.

Further, participants expressed concerns about how one could learn more by reading the policy when information is provided verbally. Thus, we postulate that new interactive techniques are needed that allow users to ask questions and seek clarifications about voice-based permission requests. First approaches, such as privacy bots (Harkous et al., 2016) that allow users to ask privacy-related questions, exist but need to be further explored in the context of Large Language Models (Leschanowsky et al., 2025; Freiberger et al., 2025). Such tools could enhance interactivity and improve accessibility of complex permission requests in VUIs. While our study used three distinct intrusiveness levels, small variations in factors such as data type, purpose or third-party sharing, can influence user perceptions. To better understand the impact of each factor on peoples' perceptions, future work should explore more fine-grained permission requests, for instance via the use of vignette studies (Abdi et al., 2021).

Finally, the linguistic structure of the decision prompts in voice permission requests is not standardized. While Amazon uses "You can say 'I approve' or 'no'", we opted for the "Please choose Decline or Accept" to ensure consistency in word length and order. Yet, prior work shows that linguistic variations can impact users' privacy decision-making (Leschanowsky et al., 2023). Moreover, response order effects can influence decision-making differently in text and voice formats (Krosnick and Alwin, 1987), but such subtle variations in word choice and order remain underexplored in the context of VUIs. Future work should explore these factors to support the development of standardized, user-friendly and non-deceptive voice prompts.

6.2. Habituation effects and voice-based permission requests salience

While we showed that participants were less likely to accept voice-based permission requests (see Table 2), habituation with the VUI may make users accustomed to such requests. In the long term, users may be less likely to critically evaluate their content as the novelty of these requests may fade over time. Initially, users may be intrigued by this new way of receiving information. However, as the novelty wears off,

[16] https://developer.amazon.com/en-US/docs/alexa/custom-skills/use-voice-forward-consent.html.

habituation effects can set in that can decrease user attention and increase acceptance (Karegar et al., 2020; Böhme and Köpsell, 2010; Vance et al., 2019). Moreover, the social aspect of voice interaction can also create a false impression of trustworthiness, as users may be more inclined to trust a voice that sounds friendly or familiar (Guerreiro and Loureiro, 2023). To counteract these effects, cues that increase the salience of voice-based permission requests could be deployed.

One possible way of increasing message salience is through the use of so-called "auditory interventions" such as error pings or earcons. Such non-speech warnings have shown better user task performance and were preferred over speech-based ones for longer audio content, since the latter can interfere with concurrent speech communication (Ming et al., 2008). Such an approach has been proposed to alert podcast listeners about potential misinformation or inaccuracies in the presented material (Pathiyan Cherumanal et al., 2024). In the context of VUIs, explicit beep notifications have been explored to inform users when private data is stored or the user has granted permission to store private data and were found helpful and noticeable (Yeasmin et al., 2020). In the context of permission requests, suitable auditory alerts could be played before users are prompted to grant their permission to indicate the importance and privacy-related implications of granting consent.

Another way of emphasizing particular parts of permission requests is through different prosodic modifications, such as the introduction of pauses, decreases in speech rate, and increases in pitch. Such modifications can be used to highlight the key elements and increase their prominence. For instance, in the utterance, "Please choose Decline or Accept", introducing a pause after 'Decline' and increasing the pitch of 'or' can emphasize that the user has a choice and underline the decision-making process rather than any of the available options. Previous work indicates that sentence stress can ease comprehension of stressed words and can lower reaction time independent of a word's syntactic function (Cutler and Foss, 1977). Moreover, humans allocate most attention to word onsets that they are less likely to predict (Astheimer and Sanders, 2011), and high activation levels allow extra cognitive resources to be allocated for processing these words by the listeners (Cole et al., 2010). In addition, studies have shown that adding filler words, pauses, or even artificial tones can help people understand spoken words better (Corley and Hartsuiker, 2011). These elements are common in natural conversations but are often left out of computer-generated speech.

6.3. Increasing friction

Another approach to improve permission request processes in VUIs, is the use of performative clauses that can introduce friction. For instance, to slow down the decision-making process, the user may be required to utter a specific phrase to officially grant their permission (e.g., "I accept the presented terms") instead of simply saying "agree" or "yes". On the system side, an agent may reiterate what the user is consenting to and then ask the user again if they would like to confirm. Recent work indicates that such interventions may be effective in fostering reflection and ultimately making the decision-making process more informed (Dubiel et al., 2024a). While these design strategies have the potential to enhance reflective thinking, they may also align with legal requirements. For example, under GDPR, "explicit consent" is required in certain situations, asking for an expressed statement of consent (European Data Protection Board (EDPB), 2020). However, the design of these strategies must be carefully considered, since slow-down nudges that have been effective in the context of social media platforms may not directly translate to conversational systems (Wang et al., 2013; Leschanowsky et al., 2023). In VUIs, slow-down interventions that increase friction could be counterproductive if they lead to user frustration or disengagement.

6.4. Implications of voice deceptive patterns

In Study 2, we showed that while subtle changes in speech rate can affect users' decision-making, the direction of the effect was not as expected. This raises questions about the applicability of voice deceptive patterns and their limitations. Further research is needed to understand the potential of deceptive patterns on prosodic attributes such as speech rate and pitch. As discussed in Section 2, these features can impact perception and appearance. It is therefore crucial to understand the impact of these features in the context of privacy permission requests and to investigate trade-offs between the likeability of voices and the interests of VUI providers and users' autonomy in making privacy decisions.

7. Limitations

We acknowledge that our study is subject to several limitations. Firstly, we only experimented with female voices. This decision was motivated by the fact that the majority of conversational agents use female voices by default, and thus could be more familiar to the users. Secondly, in our experiments, we asked users to provide their input via text rather than speaking aloud. This decision was taken to avoid privacy considerations relating to providing voice input. Thirdly, our findings are only valid for our specifically designed manipulations of speech rate. While we have discussed our design process in detail to ensure both naturalness and perceptibility of the changes, future research should explore a broader range of speech rate adjustments and investigate additional prosodic features, such as pitch variance and speech rhythm, which may also affect user decision-making. Moreover, our survey was designed to playback the voice-based permission request but asked users to make decisions via button clicks ("Accept" and "Decline"). Participants could replay the audio without restriction and were not subjected to time constraints different to real-world implementations of voice-forward consent. As these design choices may have influenced user perceptions and decision-making, future work should simulate fully voice-based interactions and potentially time pressure to better reflect real-life scenarios. Finally, the experiment was based on a one-off, hypothetical scenario where participants did not have to install a voice application or experience the implications of their consent decisions. Future work could investigate such patterns in more realistic settings and explore if interaction effects can persist over time.

8. Conclusion

With voice-enabled technologies being integrated into everyday lives, voice-based permission requests become an essential part of users' privacy controls in these interfaces. We conducted two pre-registered studies and one pre-study with two listening tasks to investigate the impact on users' decision-making and perception of voice-based permission requests and potential voice deceptive patterns. Specifically, we designed voice deceptive patterns based on changes in speech rate that are both perceived as natural and noticeably different by the users. We found that (1) users are more restrictive in accepting voice-based permission requests compared to text-based permission requests, and (2) voice deceptive patterns, i.e., changes in speech rate, can influence users' choices, however, not as originally predicted.

CRediT authorship contribution statement

Anna Leschanowsky: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Anastasia Sergeeva:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Judith Bauer:** Writing – review & editing, Writing – original draft, Visualization, Software, Project

administration, Methodology, Funding acquisition, Conceptualization. **Sheetal Vijapurapu:** Writing – original draft, Visualization, Software, Investigation, Formal analysis. **Mateusz Dubiel:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially supported by the Free State of Bavaria in the DSAI project, Germany [grant number RMF-SG20-3410-2-15-4], by the Fraunhofer-Zukunftsstiftung, Germany and by the European Union under Grant Agreement No. 101092861. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b215dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG), Germany [grant number 440719683].

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijhcs.2025.103590.

Data availability

Data is made available in the supplementary material.

References

Abdi, N., Zhan, X., Ramokapane, K.M., Such, J., 2021. Privacy norms for smart home personal assistants. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–14.

Abrokwa, D., Das, S., Akgul, O., Mazurek, M.L., 2021. Comparing security and privacy attitudes among U.S. users of different smartphone and smart-speaker platforms. In: Seventeenth Symposium on Usable Privacy and Security. SOUPS 2021, USENIX Association, pp. 139–158, URL: https://www.usenix.org/conference/soups2021/presentation/abrokwa.

Ahmad, I., Akter, T., Buher, Z., Farzan, R., Kapadia, A., Lee, A.J., 2022. Tangible privacy for smart voice assistants: Bystanders' perceptions of physical device controls. Proc. ACM Hum. Comput. Interact. 6 (CSCW2), http://dx.doi.org/10.1145/3555089.

Alexa Developer, 2023. The new Alexa design guide helps developers design skills that keep users coming back for more. URL: https://developer.amazon.com/en-US/blogs/alexa/alexa-skills-kit/2023/03/alexa-design-guide-march-2023.

Altman, I., 1975. The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding. Brooks/Cole Publishing Company.

Ammari, T., Kaye, J., Tsai, J.Y., Bentley, F., 2019. Music, search, and IoT: How people (really) use voice assistants. ACM Trans. Comput. Hum. Interact. (TOCHI) 26 (3), 1–28.

Apple, W., Streeter, L.A., Krauss, R.M., 1979. Effects of pitch and speech rate on personal attributions. J. Pers. Soc. Psychol. 37 (5), 715.

Astheimer, L.B., Sanders, L.D., 2011. Predictability affects early perceptual processing of word onsets in continuous speech. Neuropsychol. 49 (12), 3512–3516.

Baddeley, A., 1992. Working memory. Sci. 255 (5044), 556–559.

Bakhturina, E., Lavrukhin, V., Ginsburg, B., Zhang, Y., 2021. Hi-Fi multi-speaker english TTS dataset. In: Proceedings of the Annual Conference of the International Speech Communication Association. Interspeech, Brno, Czech Republic.

Baumann, T., 2017. Large-scale speaker ranking from crowdsourced pairwise listener ratings. In: Interspeech 2017. pp. 2262–2266. http://dx.doi.org/10.21437/Interspeech.2017-1697.

Belin, P., Bestelmeyer, P.E., Latinus, M., Watson, R., 2011. Understanding voice perception. Br. J. Psychol. 102 (4), 711–725.

Bem, S.L., 1981. Gender schema theory: A cognitive account of sex typing. Psychol Rev 88 (4), 354.

Bernritter, S.F., Verlegh, P.W., Smit, E.G., 2016. Why nonprofits are easier to endorse on social media: The roles of warmth and brand symbolism. J. Interact. Mark. 33 (1), 27–42.

Berry, D.C., Butler, L.T., de Rosis, F., 2005. Evaluating a realistic agent in an advice-giving task. Int. J. Hum.-Comput. Stud. 63 (3), 304–327. http://dx.doi.org/10.1016/j.ijhcs.2005.03.006.

Beyer, H., Holtzblatt, K., 1997. Contextual Design: Defining Customer-Centered Systems. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA.

Bhatia, J., Breaux, T.D., 2017. A data purpose case study of privacy policies. In: 2017 IEEE 25th International Requirements Engineering Conference. RE, pp. 394–399. http://dx.doi.org/10.1109/RE.2017.56.

Böhme, R., Köpsell, S., 2010. Trained to accept? a field experiment on consent dialogs. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '10, Association for Computing Machinery, New York, NY, USA, pp. 2403–2406. http://dx.doi.org/10.1145/1753326.1753689.

Bongard-Blanchy, K., Rossi, A., Rivas, S., Doublet, S., Koenig, V., Lenzini, G., 2021. " I am definitely manipulated, even when I am aware of it. It's ridiculous!"- Dark patterns from the end-user perspective. In: Designing Interactive Systems Conference 2021. pp. 763–776.

Bösch, C., Erb, B., Kargl, F., Kopp, H., Pfattheicher, S., 2016. Tales from the dark side: Privacy dark strategies and privacy dark patterns. Proc. Priv. Enhanc. Technol. 2016 (4), 237–254.

Brüggemeier, B., Lalone, P., 2022. Perceptions and reactions to conversational privacy initiated by a conversational user interface. Comput. Speech Lang. 71, 101269.

Chattopadhyay, A., Dahl, D.W., Ritchie, R.J., Shahin, K.N., 2003. Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising. J. Consum. Psychol. 13 (3), 198–204.

Cheng, L., Wilson, C., Liao, S., Young, J., Dong, D., Hu, H., 2020. Dangerous skills got certified: Measuring the trustworthiness of skill certification in voice personal assistant platforms. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. pp. 1699–1716.

Clark, J., Yallop, C., 1996. An Introduction to Phonetics and Phonology., Second Ed..

Cole, J., Mo, Y., Hasegawa-Johnson, M., 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. Lab. Phonol. 1 (2), 425–452.

Corley, M., Hartsuiker, R.J., 2011. Why um helps auditory word recognition: The temporal delay hypothesis. PLoS One 6 (5), e19792.

Cutler, A., Foss, D.J., 1977. On the role of sentence stress in sentence processing. Lang. Speech 20 (1), 1–10.

Data Protection Working Party, 2013. Opinion 02/2013 on apps on smart devices.

De Conca, S., 2023. The present looks nothing like the jetsons: Deceptive design in virtual assistants and the protection of the rights of users. Comput. Law Secur. Rev. 51, 105866.

Desai, S., Dubiel, M., Leiva, L.A., 2024. Examining humanness as a metaphor to design voice user interfaces. In: Proceedings of the 6th ACM Conference on Conversational User Interfaces. pp. 1–15.

Dowding, S., Gutwin, C., Cockburn, A., 2024. User speech rates and preferences for system speech rates. Int. J. Hum.-Comput. Stud. 184, 103222.

Dubiel, M., Halvey, M., Gallegos, P.O., King, S., 2020. Persuasive synthetic speech: Voice perception and user behaviour. In: Proceedings of the 2nd Conference on Conversational User Interfaces. pp. 1–9.

Dubiel, M., Leiva, L.A., Bongard-Blanchy, K., Sergeeva, A., 2024a. "Hey genie, you got me thinking about my menu choices!" Impact of proactive feedback on user perception and reflection in decision-making tasks. ACM Trans. Comput. Hum. Interact..

Dubiel, M., Sergeeva, A., Leiva, L.A., 2024b. Impact of voice fidelity on decision making: A potential dark pattern? In: Proceedings of the 29th International Conference on Intelligent User Interfaces. pp. 181–194.

Dula, E., Rosero, A., Phillips, E., 2023. Identifying dark patterns in social robot behavior. In: 2023 Systems and Information Engineering Design Symposium. SIEDS, IEEE, pp. 7–12.

Edu, J.S., Ferrer-Aran, X., Such, J., Suarez-Tangil, G., 2021. SkillVet: automated traceability analysis of Amazon Alexa skills. IEEE Trans. Dependable Secur. Comput. 20 (1), 161–175.

Edu, J., Ferrer-Aran, X., Such, J., Suarez-Tangil, G., 2022. Measuring Alexa skill privacy practices across three years. In: Proceedings of the ACM Web Conference 2022. pp. 670–680.

Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Skerry-Ryan, R., Wu, Y., 2021. Parallel tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling. In: Proceedings of the Annual Conference of the International Speech Communication Association. Interspeech, pp. 141–145. http://dx.doi.org/10.21437/Interspeech.2021-1461.

Emami-Naeini, P., Dheenadhayalan, J., Agarwal, Y., Cranor, L.F., 2021. Which privacy and security attributes most impact consumers' risk perception and willingness to purchase IoT devices? In: 2021 IEEE Symposium on Security and Privacy. SP, IEEE, pp. 519–536.

Ernst, C.-P.H., Herm-Stapelberg, N., 2020. The impact of gender stereotyping on the perceived likability of virtual assistants. In: AMCIS.

Eskimez, S.E., Wang, X., Thakker, M., Li, C., Tsai, C.-H., Xiao, Z., Yang, H., Zhu, Z., Tang, M., Tan, X., et al., 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In: 2024 IEEE Spoken Language Technology Workshop. SLT, IEEE, pp. 682–689.

European Commission, 2016. Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation) (text with EEA relevance). URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

European Data Protection Board (EDPB), 2020. Guidelines 05/2020 on consent under regulation 2016/679.

European Data Protection Board (EDPB), 2023. Guidelines 03/2022 on deceptive design patterns in social media platform interfaces: how to recognise and avoid them.

European Parliament and Council of the European Union, 2022a. Regulation (EU) 2022/1925 of the European parliament and of the council of 14 september 2022 on contestable and fair markets in the digital sector and amending directives (EU) 2019/1937 and (EU) 2020/1828 (digital markets act).

European Parliament and Council of the European Union, 2022b. Regulation (EU) 2022/2065 of the European parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/EC (digital services act).

European Parliament and Council of the European Union, 2024. Regulation (EU) 2024/1689 of the European parliament and of the council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (artificial intelligence act).

Finstad, K., 2010. Response interpolation and scale sensitivity: Evidence against 5-point scales. J. Usability Stud. 5 (3), 104–110.

Freiberger, V., Fleig, A., Buchmann, E., 2025. "You don't need a university degree to comprehend data protection this way": LLM-powered interactive privacy policy assessment. In: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. In: CHI EA '25, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3706599.3719816.

Fruchter, N., Liccardi, I., 2018. Consumer attitudes towards privacy and security in home assistants. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–6.

Gaiser, F., Utz, S., 2023. Is hearing really believing? The importance of modality for perceived message credibility during information search with smart speakers. J. Media Psychol. Theor. Methods Appl..

Geipel, J., Hadjichristidis, C., Savadori, L., Keysar, B., 2023. Language modality influences risk perception: Innovations read well but sound even better. Risk Anal. 43 (3), 558–570.

Goodman, K.L., Mayhorn, C.B., 2023. It's not what you say but how you say it: Examining the influence of perceived voice assistant gender and pitch on trust and reliance. Appl. Ergon. 106, 103864.

Gray, C.M., Chivukula, S.S., Lee, A., 2020. What kind of work do" asshole designers" create? Describing properties of ethical concern on reddit. In: Proceedings of the 2020 Acm Designing Interactive Systems Conference. pp. 61–73.

Gray, C.M., Kou, Y., Battles, B., Hoggatt, J., Toombs, A.L., 2018. The dark (patterns) side of UX design. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–14.

Gray, C.M., Santos, C., Bielova, N., Toth, M., Clifford, D., 2021. Dark patterns and the legal requirements of consent banners: An interaction criticism perspective. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–18.

Groß, T., 2023. Toward valid and reliable privacy concern scales: The example of iuipc-8. In: Human Factors in Privacy Research. Springer International Publishing Cham, pp. 55–81.

Guerreiro, J., Loureiro, S.M.C., 2023. I am attracted to my cool smart assistant! Analyzing attachment-aversion in AI-human relationships. J. Bus. Res. 161, 113863.

Gunawan, J., Santos, C., Kamara, I., 2022. Redress for dark patterns privacy harms? A case study on consent interactions. In: Proceedings of the 2022 Symposium on Computer Science and Law. pp. 181–194.

Hall, J.A., 1978. Gender effects in decoding nonverbal cues. Psychol. Bull. 85 (4), 845.

Hall, J.A., Carter, J.D., Horgan, T.G., 2000. Gender differences in nonverbal communication of emotion. Gend. Emot. Soc. Psychol. Perspect. 97–117.

Hamilton, K., Shih, S.-I., Mohammed, S., 2016. The development and validation of the rational and intuitive decision styles scale. J. Pers. Assess. 98 (5), 523–535.

Harkous, H., Fawaz, K., Shin, K.G., Aberer, K., 2016. PriBots: Conversational privacy with chatbots. In: Twelfth Symposium on Usable Privacy and Security. SOUPS 2016.

Holbrook, A.L., Krosnick, J.A., Moore, D., Tourangeau, R., 2007. Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. Public Opin. Q. 71 (3), 325–348. http://dx.doi.org/10.1093/poq/nfm024.

Holmberg, E.B., Hillman, R.E., Perkell, J.S., 1988. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. J. Acoust. Soc. Am. 84 (2), 511–529.

Huang, R., Lam, M.W.Y., Wang, J., Su, D., Yu, D., Ren, Y., Zhao, Z., 2022. FastDiff: A fast conditional diffusion model for high-quality speech synthesis. In: Raedt, L.D. (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. IJCAI, pp. 4157–4163.

International Telecommunication Union, 2015. Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems. URL: https://www.itu.int/rec/R-REC-BS.1534/en.

Ischen, C., Araujo, T.B., Voorveld, H.A., Van Noort, G., Smit, E.G., 2022. Is voice really persuasive? The influence of modality in virtual assistant interactions and two alternative explanations. Internet Res. 32 (7), 402–425.

Joly, A., Nicolis, M., Peterova, E., Lombardi, A., Abbas, A., van Korlaar, A., Hussain, A., Sharma, P., Moinet, A., Lajszczak, M., Karanasou, P., Bonafonte, A., Drugman, T., Sokolova, E., 2023. Controllable emphasis with zero data for text-to-speech. In: Proceedings of the ISCA Workshop on Speech Synthesis. SSW, Grenoble, France, pp. 113–119. http://dx.doi.org/10.21437/SSW.2023-18.

Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., Wu, Z., Qin, T., Li, X.-Y., Ye, W., Zhang, S., Bian, J., He, L., Li, J., Zhao, S., 2024. NaturalSpeech 3: zero-shot speech synthesis with factorized codec and diffusion models. In: Proceedings of the 41st International Conference on Machine Learning. ICML '24, JMLR.org.

Jungers, M.K., Hupp, J.M., 2009. Speech priming: Evidence for rate persistence in unscripted speech. Lang. Cogn. Process. 24 (4), 611–624.

Kang, H., Oh, J., 2023. Communication privacy management for smart speaker use: Integrating the role of privacy self-efficacy and the multidimensional view. New Media Soc. 25 (5), 1153–1175. http://dx.doi.org/10.1177/14614448211026611.

Karegar, F., Pettersson, J.S., Fischer-Hübner, S., 2020. The dilemma of user engagement in privacy notices: Effects of interaction modes and habituation on user attention. ACM Trans. Priv. Secur. 23 (1), http://dx.doi.org/10.1145/3372296.

Kawakita, J., 1982. The Original KJ Method. Kawakita Research Institute, Tokyo.

Kiesel, A., Wagener, A., Kunde, W., Hoffmann, J., Fallgatter, A.J., Stöcker, C., 2006. Unconscious manipulation of free choice in humans. Conscious. Cogn. 15 (2), 397–408.

Kochanski, G., Grabe, E., Coleman, J., Rosner, B., 2005. Loudness predicts prominence: Fundamental frequency lends little. J. Acoust. Soc. Am. 118 (2), 1038–1054.

Korff, S., Böhme, R., 2014. Too much choice: End-user privacy decisions in the context of choice proliferation. In: 10th Symposium on Usable Privacy and Security. SOUPS 2014, pp. 69–87.

Krisam, C., Dietmann, H., Volkamer, M., Kulyk, O., 2021. Dark patterns in the wild: Review of cookie disclaimer designs on top 500 german websites. In: Proceedings of the 2021 European Symposium on Usable Security. pp. 1–8.

Krosnick, J.A., Alwin, D.F., 1987. An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opin. Q. 51 (2), 201–219.

Kyi, L., Ammanaghatta Shivakumar, S., Santos, C.T., Roesner, F., Zufall, F., Biega, A.J., 2023. Investigating deceptive design in gdpr's legitimate interest. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–16.

Lau, J., Zimmerman, B., Schaub, F., 2018. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. Proc. ACM Hum. Comput. Interact. 2 (CSCW), 1–31.

Lentzsch, C., Shah, S.J., Andow, B., Degeling, M., Das, A., Enck, W., 2021. Hey Alexa, is this skill safe?: Taking a closer look at the Alexa skill ecosystem. In: Network and Distributed Systems Security (NDSS) Symposium.

Leschanowsky, A., Popp, B., Peters, N., 2023. Privacy strategies for conversational AI and their influence on users' perceptions and decision-making. In: Proceedings of the 2023 European Symposium on Usable Security. pp. 296–311.

Leschanowsky, A., Rech, S., Popp, B., Bäckström, T., 2024. Evaluating privacy, security, and trust perceptions in conversational AI: A systematic review. Comput. Hum. Behav. 159, 108344. http://dx.doi.org/10.1016/j.chb.2024.108344.

Leschanowsky, A., Salamatjoo, F., Kolagar, Z., Popp, B., 2025. Expert-generated privacy Q&A dataset for conversational AI and user study insights. In: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. In: CHI EA '25, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3706599.3720014, URL: https://doi.org/10.1145/3706599.3720014.

Levitan, S.I., Maredia, A., Hirschberg, J., 2018. Acoustic-prosodic indicators of deception and trust in interview dialogues. In: Interspeech. pp. 416–420.

Liao, S., Cheng, L., Cai, H., Guo, L., Hu, H., 2023. Skillscanner: Detecting policy-violating voice applications through static analysis at the development phase. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. pp. 2321–2335.

Liao, S., Wilson, C., Cheng, L., Hu, H., Deng, H., 2020. Measuring the effectiveness of privacy policies for voice assistant applications. In: Proceedings of the 36th Annual Computer Security Applications Conference. pp. 856–869.

Lin, V.Z., Parkin, S., 2020. Transferability of privacy-related behaviours to shared smart home assistant devices. In: 2020 7th International Conference on Internet of Things: Systems, Management and Security. IOTSMS, pp. 1–8. http://dx.doi.org/10.1109/IOTSMS52051.2020.9340199.

Luguri, J., Strahilevitz, L.J., 2021. Shining a light on dark patterns. J. Leg. Anal. 13 (1), 43–109.

Malkin, N., Wagner, D., Egelman, S., 2022. Runtime permissions for privacy in proactive intelligent assistants. In: Eighteenth Symposium on Usable Privacy and Security. SOUPS 2022, pp. 633–651.

Masotina, M., Spagnolli, A., 2022. Transparency of privacy notices and contextualisation: effectively conveying information without words. Behav. Inf. Technol. 41 (10), 2120–2150.

Mathur, A., Kshirsagar, M., Mayer, J., 2021. What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–18.

Mazanec, N., Mccall, G.J., 1976. Sex factors and allocation of attention in observing persons. J. Psychol. 93 (2), 175–180.

Mildner, T., Cooney, O., Meck, A.-M., Bartl, M., Savino, G.-L., Doyle, P.R., Garaialde, D., Clark, L., Sloan, J., Wenig, N., et al., 2024. Listening to the voices: Describing ethical caveats of conversational user interfaces according to experts and frequent users. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–18.

Ming, L.J., Aziz, F.A., Sahari, B., 2008. A study on real-time auditory feedback technique in manipulation task. In: 2008 International Symposium on Information Technology. Vol. 1, IEEE, pp. 1–6.

Molden, D.C., 2014. Understanding priming effects in social psychology: What is "social priming" and how does it occur? Soc. Cogn. 32 (Supplement), 1–11.

Morel, V., Pardo, R., 2020. Sok: Three facets of privacy policies. In: Proceedings of the 19th Workshop on Privacy in the Electronic Society. WPES '20, Association for Computing Machinery, New York, NY, USA, pp. 41–56. http://dx.doi.org/10.1145/3411497.3420216.

Mustafa, A., Pia, N., Fuchs, G., 2021. StyleMelGAN: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP, Toronto, Canada, pp. 6034–6038.

National Public Media, 2022. The smart audio report | national public media — nationalpublicmedia.com. https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/. (Accessed 05 September 2023).

Nguyen, B., Cardinaux, F., Uhlich, S., 2023. Autotts: End-to-end text-to-speech synthesis through differentiable duration modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49357.2023.10095431.

Nowak, G.J., Phelps, J.E., 1992. Understanding privacy concerns: An assessment of consumers' information-related knowledge and beliefs. J. Direct Mark. 6 (4), 28–39.

Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., Sorokowska, A., 2017. Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. Psychon. Bull. Rev. 24, 856–862.

Owens, K., Gunawan, J., Choffnes, D., Emami-Naeini, P., Kohno, T., Roesner, F., 2022. Exploring deceptive design patterns in voice interfaces. In: Proceedings of the 2022 European Symposium on Usable Security. pp. 64–78.

Pal, D., Arpnikanondt, C., Razzaque, M.A., 2020. Personal information disclosure via voice assistants: The personalization–Privacy paradox. SN Comput. Sci. 1 (5), http://dx.doi.org/10.1007/s42979-020-00287-9.

Pathiyan Cherumanal, S., Gadiraju, U., Spina, D., 2024. Everything we hear: Towards tackling misinformation in podcasts. In: Proceedings of the 26th International Conference on Multimodal Interaction. pp. 596–601.

Payne, B.K., Brown-Iannuzzi, J.L., Loersch, C., 2016. Replicable effects of primes on human behavior. J. Exp. Psychol. [Gen.] 145 (10), 1269.

Pias, S.B.H., Huang, R., Williamson, D.S., Kim, M., Kapadia, A., 2024. The impact of perceived tone, age, and gender on voice assistant persuasiveness in the context of product recommendations. In: Proceedings of the 6th ACM Conference on Conversational User Interfaces. pp. 1–15.

Potoglou, D., Dunkerley, F., Patil, S., Robinson, N., 2017. Public preferences for internet surveillance, data retention and privacy enhancing services: Evidence from a pan-European study. Comput. Hum. Behav. 75, 811–825.

Puts, D.A., Gaulin, S.J., Verdolini, K., 2006. Dominance and the evolution of sexual dimorphism in human voice pitch. Evol. Hum. Behav. 27 (4), 283–296.

Puts, D.A., Hill, A.K., Bailey, D.H., Walker, R.S., Rendall, D., Wheatley, J.R., Welling, L.L., Dawood, K., Cárdenas, R., Burriss, R.P., et al., 2016. Sexual selection on male vocal fundamental frequency in humans and other anthropoids. Proc. R. Soc. Biol. Sci. 283 (1829), 20152830.

Raitio, T., Li, J., Seshadri, S., 2022. Hierarchical prosody modeling and control in non-autoregressive parallel neural TTS. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP, Singapore, pp. 7587–7591. http://dx.doi.org/10.1109/ICASSP43922.2022.9746253.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T., 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In: Proceedings of the International Conference on Learning Representations. ICLR, virtual, Austria.

Renz, A., Neff, T., Baldauf, M., Maier, E., 2023. Authentication methods for voice services on smart speakers–a multi-method study on perceived security and ease of use. I-Com 22 (1), 67–81.

Roca, I., Johnson, W., 1999. A Course in Phonology. Blackwell Publishers, Ltd., Oxford.

Rodero, E., 2016. Influence of speech rate and information density on recognition: The moderate dynamic mechanism. Media Psychol. 19 (2), 224–242.

Rodero, E., Larrea, O., Rodríguez-de Dios, I., Lucas, I., 2022. The expressive balance effect: perception and physiological responses of prosody and gestures. J. Lang. Soc. Psychol. 41 (6), 659–684.

Rzepka, C., Berger, B., Hess, T., 2022. Voice assistant vs. Chatbot–examining the fit between conversational agents' interaction modalities and information search tasks. Inf. Syst. Front. 24 (3), 839–856.

Sandhu, R., Dyson, B.J., 2012. Re-evaluating visual and auditory dominance through modality switching costs and congruency analyses. Acta Psychol. 140 (2), 111–118.

Schäfer, C., McCarthy, O., contributors, 2020. ForwardTacotron. https://github.com/as-ideas/ForwardTacotron.

Schaub, F., Balebako, R., Cranor, L.F., 2017. Designing effective privacy notices and controls. IEEE Internet Comput. 21 (3), 70–77. http://dx.doi.org/10.1109/MIC.2017.75.

Schaub, F., Balebako, R., Durity, A.L., Cranor, L.F., 2015. A design space for effective privacy notices. In: Eleventh Symposium on Usable Privacy and Security. SOUPS 2015, pp. 1–17.

Schild, C., Stern, J., Zettler, I., 2020. Linking men's voice pitch to actual and perceived trustworthiness across domains. Behav. Ecol. 31 (1), 164–175.

Schirmer, A., Chiu, M.H., Lo, C., Feng, Y.-J., Penney, T.B., 2020. Angry, old, male–and trustworthy? How expressive and person voice characteristics shape listener trust. PLoS One 15 (5), e0232431.

Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., Herre, J., 2018. webMUSHRA—A comprehensive framework for web-based listening tests.

Schomakers, E.-M., Lidynia, C., Müllmann, D., Ziefle, M., 2019. Internet users' perceptions of information sensitivity–insights from Germany. Int. J. Inf. Manage. 46, 142–150.

Schuetzler, R.M., Grimes, G.M., Giboney, J.S., Nunamaker, Jr., J.F., 2018. The influence of conversational agents on socially desirable responding. In: Proceedings of the 51st Hawaii International Conference on System Sciences. p. 283.

Seshadri, S., Raitio, T., Castellani, D., Li, J., 2022. Emphasis control for parallel neural TTS. In: Proceedings of the Annual Conference of the International Speech Communication Association. Interspeech, Incheon, Korea, pp. 3378–3382. http://dx.doi.org/10.21437/Interspeech.2022-411.

Seymour, W., Abdi, N., Ramokapane, K.M., Edu, J., Suarez-Tangil, G., Such, J., 2024. Voice app developer experiences with Alexa and google assistant: Juggling risks, liability, and security. In: 33rd USENIX Security Symposium. USENIX Security 24, USENIX Association, Philadelphia, PA, pp. 5035–5052.

Seymour, W., Binns, R., Slovak, P., Van Kleek, M., Shadbolt, N., 2020. Strangers in the room: Unpacking perceptions of 'smartness' and related ethical concerns in the home. In: Proceedings of the 2020 ACM Designing Interactive Systems Conference. DIS '20, Association for Computing Machinery, New York, NY, USA, pp. 841–854. http://dx.doi.org/10.1145/3357236.3395501.

Seymour, W., Cote, M., Such, J., 2022. Can you meaningfully consent in eight seconds? Identifying ethical issues with verbal consent for voice assistants. In: Proceedings of the 4th Conference on Conversational User Interfaces. CUI '22, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3543829.3544521.

Seymour, W., Cote, M., Such, J., 2023. Legal obligation and ethical best practice: Towards meaningful verbal consent for voice assistants. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3544548.3580967.

Shechtman, S., Fernandez, R., Haws, D., 2021. Supervised and unsupervised approaches for controlling narrow lexical focus in sequence-to-sequence speech synthesis. In: Proceedings of the IEEE Spoken Language Technology Workshop. SLT, Shenzhen, China, pp. 431–437. http://dx.doi.org/10.1109/SLT48900.2021.9383591.

Shibuya, T., Takida, Y., Mitsufuji, Y., 2024. BIGVSAN: Enhancing gan-based neural vocoders with slicing adversarial network. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP, pp. 10121–10125. http://dx.doi.org/10.1109/ICASSP48485.2024.10446121.

Shih, F., Liccardi, I., Weitzner, D., 2015. Privacy tipping points in smartphones privacy preferences. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 807–816.

Simpson, A.P., 2009. Phonetic differences between male and female speech. Lang. Linguist. Compass 3 (2), 621–640.

Smith, H.J., Dinev, T., Xu, H., 2011. Information privacy research: an interdisciplinary review. MIS Q. 989–1015.

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., et al., 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. IEEE Trans. Pattern Anal. Mach. Intell..

Taylor, P., 2009. Text-To-Speech Synthesis. Cambridge University Press.

Titze, I.R., Martin, D.W., 1998. Principles of voice production.

Tooley, K.M., Konopka, A.E., Watson, D.G., 2018. Assessing priming for prosodic representations: Speaking rate, intonational phrase boundaries, and pitch accenting. Mem. Cogn. 46, 625–641.

Tsao, Y.-C., Weismer, G., Iqbal, K., 2006. Interspeaker variation in habitual speaking rate: Additional evidence.

Tuncer, R., Sergeeva, A., Bongard-Blanchy, K., Distler, V., Doublet, S., Koenig, V., 2023. Running out of time (rs): effects of scarcity cues on perceived task load, perceived benevolence and user experience on e-commerce sites. Behav. Inf. Technol. 1–19.

Utz, C., Degeling, M., Fahl, S., Schaub, F., Holz, T., 2019. (Un) informed consent: Studying GDPR consent notices in the field. In: Proceedings of the 2019 Acm Sigsac Conference on Computer and Communications Security. pp. 973–990.

Valoggia, P., Sergeeva, A., Rossi, A., Botes, M., 2024. Learning from the dark side about how (not) to engineer privacy: Analysis of dark patterns taxonomies from an ISO 29100 perspective. In: Proceedings of the 10th International Conference on Information Systems Security and Privacy - ICISSP. INSTICC. SciTePress, pp. 774–784. http://dx.doi.org/10.5220/0012393100003648.

Vance, A., Eargle, D., Jenkins, J.L., Kirwan, C.B., Anderson, B.B., 2019. The fog of warnings: How non-essential notifications blur with security warnings. In: Fifteenth Symposium on Usable Privacy and Security. SOUPS 2019, USENIX Association, Santa Clara, CA, pp. 407–420.

Varghese, A.L., Nilsen, E.S., 2020. Is that how you should talk to her? Using appropriate prosody affects adults', but not children's, judgments of communicators' competence. J. Lang. Soc. Psychol. 39 (5–6), 738–750.

Vixen Labs, 2023. AI consumer index 2023.

Vukovic, J., Jones, B.C., Feinberg, D.R., DeBruine, L.M., Smith, F.G., Welling, L.L., Little, A.C., 2011. Variation in perceptions of physical dominance and trustworthiness predicts individual differences in the effect of relationship context on women's preferences for masculine pitch in men's voices. Br. J. Psychol. 102 (1), 37–48.

Wambsganss, T., Zierau, N., Söllner, M., Käser, T., Koedinger, K.R., Leimeister, J.M., 2022. Designing conversational evaluation tools: A comparison of text and voice modalities to improve response quality in course evaluations. Proc. ACM Hum. Comput Interact. 6 (CSCW2), 1–27.

Wang, Y., Leon, P.G., Scott, K., Chen, X., Acquisti, A., Cranor, L.F., 2013. Privacy nudges for social media: an exploratory facebook study. In: Proceedings of the 22nd International Conference on World Wide Web. In: WWW '13 Companion, Association for Computing Machinery, New York, NY, USA, pp. 763–770. http://dx.doi.org/10.1145/2487788.2488038.

Westin, A.F., 1967. Privacy and Freedom. Atheneum, New York.

Wottrich, V.M., van Reijmersdal, E.A., Smit, E.G., 2018. The privacy trade-off for mobile app downloads: The roles of app value, intrusiveness, and privacy concerns. Decis. Support Syst. 106, 44–52.

Yeasmin, F., Das, S., Bäckström, T., 2020. Privacy analysis of voice user interfaces. In: Conference of Open Innovations Association, FRUCT. Vol. 6.

Zalkow, F., Sani, P., Fast, M., Bauer, J., Joshaghani, M., Kayyar, K., Habets, E.A.P., Dittmar, C., 2023. The AudioLabs system for the blizzard challenge 2023. In: Proceedings of the Blizzard Challenge Workshop. Grenoble, France, pp. 63–68.