Computer Speech and Language 96 (2026) 101884

Contents lists available at ScienceDirect

Computer Speech & Language

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

# Public perceptions of speech technology trust in the United Kingdom[☆]

Jennifer Williams [a,*,1], Tayyaba Azim [a,1], Anna-Maria Piskopani [b], Richard Hyde [b], Shuo Zhang [c], Zack Hodari [d]

[a] *University of Southampton, Electronics and Computer Science, Southampton, UK*
[b] *University of Nottingham, Horizon Institute of Digital Economy, Nottingham, UK*
[c] *Bose Corporation, Audio Systems Machine Learning, Boston, MA, USA*
[d] *Papercup Technologies Ltd., Research and Development, London, UK*

## ARTICLE INFO

*Keywords:*
Speech privacy
Speech security
Speech technology regulation
Survey
Public perception
Trust
Privacy rights

## ABSTRACT

Speech technology is now pervasive throughout the world, impacting a variety of socio-technical use-cases. Speech technology is a broad term encompassing capabilities that translate, analyse, transcribe, generate, modify, enhance, or summarise human speech. Many of the technical features and the possibility of speech data misuse are not often revealed to the users of such systems. When combined with the rapid development of AI and the plethora of use-cases where speech-based AI systems are now being applied, the consequence is that researchers, regulators, designers and government policymakers still have little understanding of the public's perception of speech technology. Our research explores the public's perceptions of trust in speech technology by asking people about their experiences, awareness of their rights, their susceptibility to being harmed, their expected behaviour, and ethical choices governing behavioural responsibility. We adopt a multidisciplinary lens to our work, in order to present a fuller picture of the United Kingdom (UK) public perspective through a series of socio-technical scenarios in a large-scale survey. We analysed survey responses from 1,000 participants from the UK, where people from different walks of life were asked to reflect on existing, emerging, and hypothetical speech technologies. Our socio-technical scenarios are designed to provoke and stimulate debate and discussion on principles of trust, privacy, responsibility, fairness, and transparency. We found that gender is a statistically significant factor correlated to awareness of rights and trust. We also found that awareness of rights is statistically correlated to perceptions of trust and responsible use of speech technology. By understanding the notions of responsibility in behaviour and differing perspectives of trust, our work encapsulates the current state of public acceptance of speech technology in the UK. Such an understanding has the potential to affect how regulatory and policy frameworks are developed, how the UK invests in its AI research and development ecosystem, and how speech technology that is developed within the UK might be received by global stakeholders.

[☆] This article is part of a Special issue entitled: 'Security and Privacy in Speech Communication' published in Computer Speech & Language.

[*] Corresponding author.

*E-mail addresses:* j.williams@soton.ac.uk (J. Williams), ta7g21@soton.ac.uk (T. Azim), anna-maria.piskopani@nottingham.ac.uk (A.-M. Piskopani), richard.hyde@nottingham.ac.uk (R. Hyde), shuo_zhang@bose.com (S. Zhang), zack@papercup.com (Z. Hodari).

[1] Equal contribution.

https://doi.org/10.1016/j.csl.2025.101884
Received 5 August 2024; Received in revised form 23 August 2025; Accepted 27 August 2025

Available online 12 September 2025

## 1. Introduction

Speech technology is meeting at the crossroads where interdisciplinary work is not only becoming more common but it is also becoming necessary in order to address complex emerging challenges in society. The potential for speech data misuse has recently been made more clear with ongoing efforts in the speech community to address voice privacy (Tomashenko et al., 2020). From the general public to the academic community, voice spoofing and voice deepfakes is now a growing concern, especially with voice data seemingly captured in a multitude of different contexts from devices that many people use on a daily basis. The academic community has attempted to prepare for this through activities such as the biennial international ASVspoof challenges in 2015 (Wu et al., 2015), 2017 (Kinnunen et al., 2017), 2019 (Todisco et al., 2019), 2021 (Yamagishi et al., 2021), and 2024 (Wang et al., 2024). However, these challenges are often limited by the type of datasets that are curated by the organisers, and the challenges do not always capture a wider view of the problem of voice authenticity (Müller et al., 2024). Further, many commonly-used speech devices in the home are known to malfunction due to speaker characteristics such as accent (Pal et al., 2019; Nacimiento-García et al., 2024) or edge-device listening mode (Sweeney and Davis, 2020). Products that have difficulty with certain speaker accents can reinforce negative stereotypes (Austin, 2019) at best, or result in a safety-critical situation when people rely on speech-enabled devices for emergencies (Picard et al., 2020). In addition, there are increasing concerns about how consumer data can be used or misused by corporations which is not limited to speech data, and may include pattern-of-life analysis (Veale et al., 2018). Consumers may feel frustrated if they have concerns about their voice data privacy but may also feel that they are unable to opt-out of voice technologies (Koelle et al., 2018; Datta et al., 2018) because such technologies are now ubiquitous and facilitate participation in modern and professional society.

Our study is the first to explicitly explore public attitudes of trust towards speech technologies in the United Kingdom (UK), and the first to explore the alignment of this trust with human rights. From a broad remit of different types of speech technologies, both existing and anticipated (Section 2.3–2.11), we introduce a new set of four survey variables (Section 3.7–3.10) that are grounded in human rights. We define, measure and compare (1) perceived awareness of rights, (2) perceived responsibility, (3) perceived trust, and (4) exposure to risks. One of the key strengths of our research methodology is that we introduce survey participants to technologies that are still in prototype stages and not yet commercialised. Thus we avoid having to measure trust or responsibility through simple measures of product purchasing while also future-proofing some aspects of our findings. Further, our survey variables measure socio-technical issues directly and indirectly from targeted survey questions, demonstrating a reliable process for creating and evaluating customised survey variables that will translate to survey-based research in other domains. Using these new variables, we investigate the following four socio-technical research questions, blending aspects of speech technology privacy and security:

- **RQ1**: Is appropriate and responsible *use* of speech technology influenced by the public's understanding of legal rights and human rights?
- **RQ2**: Is *trust* in speech technology affected by the public's understanding of legal rights and human rights?
- **RQ3**: Are there any correlations between trust in speech technology and the public's understanding of potential risk exposures?
- **RQ4:** Do perceptions of exposure to hazards, knowledge of human rights and the law, trust in speech and voice technology, and responsible behaviour depend on individual's demographic attributes such as age, gender and level of education?

These research questions are driven by the timely global discussion taking place on AI safety (Gyevnár and Kasirzadeh, 2025), and the UK Government's recent policy of taking a "pro-innovation approach" to AI regulation (UK Government, 2023b, 2024). The 2025 International AI Safety Report (Bengio et al., 2025) notes that voice impersonation attacks are one of the greatest risks of AI that society faces, due to how voice data is (un-)protected. One of the UK's leading communications regulators, Ofcom, found that 43% of people in the UK aged 16+ have encountered a deepfake that demeans, defrauds, or dis-informs. Ofcom further found that women journalists are often targeted most often for sexually explicit deepfakes, which can have the impact of silencing critical journalism (UK Government, 2023a). Public perceptions of risk and harm from speech data misuse could have potential negative impacts on speech technology adoption, innovation, and continued use. Our exploration of public perceptions of trust in the UK is a first step towards understanding societal concerns about speech technology in more depth, with the aim of informing future innovation cycles and public policy.

Public perceptions of privacy and security measured from surveys can be indicative for how people engage with digital technology (Liu et al., 2024). Attitudes towards artificial intelligence are highly correlated with perception of risks, and perception of risks can differ significantly across global regions and national borders (Yigitcanlar et al., 2023). Artificial intelligence is a very broad term, encompassing many aspects of autonomous systems built for various purposes, or even foundational models (Bommasani et al., 2021) that are not trained for particular applications. In this work, we are focused on the subdiscipline of artificial intelligence concerned with speech technology, which captures and/or uses human voice data.

The innovation landscape for speech technology has been shifting quickly, but it is not clear whether public trust is keeping pace. We argue that now is the time to re-assess how emerging speech technologies are perceived. Whether or not speech technologies will be trusted in the future depends on how all stakeholders come together to envision a positive future ecosystem of responsible research innovation. The challenges that speech technology presents to regulators and policymakers are socio-technical in nature, and require insight beyond traditional purely-technical thinking. Therefore, in this work, we examine public perceptions of trust in speech technology through a multidisciplinary lens including law, speech science, human–computer interaction, and artistic free expression. Our intended readership of this article spans legal scholars who may benefit from a deeper understanding of the speech technology landscape, policy makers in the UK and beyond who may be interested in linking human rights to speech technologies, and speech technology developers who are curious about the interplay between technology innovation and policy.

**Table 1**
Example scenarios based on existing/known speech technologies (see Appendix A for full set of scenarios)

| Application Domain | Existing Socio-Technical Scenarios | Relevance |
|---|---|---|
| A. Health Domain | Scenario A1 – Smart Hearing Aid: Jane is hard of hearing (almost completely deaf) and uses a special hearing aid to help her throughout the day. This hearing aid can enhance specific voices that are within 5–10 m from where she is standing. It can also generate transcripts of conversations, which she can access on her phone and read to herself | Speech Enhancement, Speech Recognition |
| B. Academia | Scenario B1 – Creating Audio Books: AI generative tools can also convert audio to text and text to audio. They could be used for artistic purposes such as creating audio books. | Speech Synthesis |
| C. Digital Voice Recording Market | Scenario C1 – Preserving Voice Memory of Loved Ones: They say that the first thing you forget about a person is the sound of their voice. Preserving people's voice, memories and personal stories can be valuable to families and next generations. Interviews with loved ones in combination with their favourite music can construct a voice and memories album. | Voice Cloning, Expressive speech synthesis |
| D. Smart Environments | Scenario D1 – Vehicle: While you are driving, you can use your smart vehicle's audio capture capabilities to create a grocery shopping list, order takeout, or buy an audio book from the car without taking your eyes off the road using a pre-installed voice assistant. Your voice assistant can also predict some of your needs and make suggestions or choose and shop for products using your prior history, lifestyle behaviours and patterns as well as your budget settings. | Task-based Agents, Natural Language Understanding |
| E. Films Industry | Scenario E1 – Translation Services: Automatic translation services provide films and television shows that are not in your native language. For example, your preferred television news service or radio station is only provided in a rare foreign language, so you utilise a translation service. The translations happen automatically and instead of providing captions, the translated content is spoken just like a person is speaking. Still, you are told that an AI algorithm has performed the translation service. | Machine Translation, Speech-to-Speech Translation |
| F. Voice Conferencing | Scenario F1 – Work: The company you work for requires you to use a video conferencing tool to engage in a high-stakes meeting with an international business partner. | Speech enhancement, Speech-to-speech Translation |

## 2. Background and motivation

In this section we motivate the need for our study (Section 2.1), provide insights about speech and audio technology legal protections (Section 2.2), followed by a descriptive overview of nine types of known speech technologies concerned with our study (Section 2.3–2.11). Participants evaluated a large set of different micro scenarios on speech technology applications. Each of the nine speech technologies is labelled with a letter, and individual scenarios indicated by an additional number. This allowed us to more easily track which technologies were used in our study. Within our set of scenarios involving audio capture by speech technologies that already exist and likely known to most consumers (designated by letters A–F), we discuss: speech enhancement (A.1, F.1); speech recognition (A.1); video understanding (A.2); audio scene description (A.2); text-to-speech synthesis (A.3, B.1, C.1); voice cloning (A.3, C.1, F.2); task-based agents and natural language understanding (D.1); machine translation (E.1); speech-to-speech translation (E.1, F.1). In Table 1, we provide an abbreviated summary of the application domains (A–F), specific scenarios from our survey, and their relevance to speech technology. Likewise, in Table 2 we present an abbreviated summary of *emerging* speech technologies (designated by letters G–I) that are not yet fully commercialised as consumer products at the time of this writing, and likely unknown to most consumers. Further full details from each of these tables are available in Appendix A Tables A.5 and A.6. These emerging and hypothetical scenarios include: speech recognition (G.1, H.2); conversational prosody (G.1, H.2, I.3); task-based agents and natural language understanding (G.1, H.2, I.3); voice cloning (G.2, I.2); audio understanding (H.1, I.3); autonomous robots (H.2). The example emerging technologies in Table 2 describe uses of speech technology imagined by the authors as products of the future that build upon current example capabilities from Table 1.

### 2.1. Motivation

Speech technology products that are developed and sold by industry companies have the potential to impact lives positively or negatively. Prior research on public perceptions indicates that despite widespread adoption of conversational voice AI systems, users express concerns regarding privacy, security and trust in these systems (Leschanowsky et al., 2024). Users of these products identified surveillance anxiety, privacy and security risks (Kowalczuk, 2018). At the same time, research shows that the convenience of using these assistants can overcome any related concerns especially when there is a lack of awareness of the extent and complexity of the

**Table 2**
Example scenarios based on imagined/hypothetical speech technologies (see Appendix A for full set of scenarios).

| Application Domain | Hypothetical Socio-Technical Scenarios | Relevance |
|---|---|---|
| G. Interactive Children's Toys | Scenario G1 – AI Dolls: Someone gives your child an interactive doll as a present. It is an internet-connected doll equipped with speech recognition systems and AI based learning features, operating as an Internet of things (IoT) device. The doll remembers the child's play history and past conversations and can suggest new games and topics. The doll is named "Lucy" and she interacts with your children and can answer their questions. | Speech Recognition, Speech Synthesis, Task-based agents and NLU, Autonomous Robots |
| H. Health Domain | Scenario H1 – Health Monitoring App: You have recently been diagnosed with a breathing disorder that requires supervision by your medical doctor. Your doctor recommends a new AI-based app that uses audio from a smartphone or similar device, allowing instant feedback about the condition quickly and can alert if there is a medical emergency or not. | |
| I. Audio Immersive Projects in Arts | Scenario I1 – Live Singing: An audio capture system is set up at a venue where you go to see a singer perform. It is an experimental performance where your voice is captured prior to the performance and then used to perform a series of songs. It's your voice and its sound that the singer uses to perform the first song - the singer sounds like you! | Voice Cloning |

risk (Lau et al., 2018). Significant advantages of using these technologies were identified by people with disabilities. For example, they identified ease of use compared to existing technology and the ability to live more independently and complete everyday tasks, as well as speech therapy, learning support, and memory support benefits (Pradhan et al., 2018; Kulkarni et al., 2022).

Recent work has begun to explore trust in general applications of AI, such as the impact of consumer goods prices on satisfactory shopping experiences with respect to recommender systems using a theoretical behavioural model of stimulus-organism-response (Shan and Li, 2025). While consumer products and customer satisfaction are important to a healthy AI ecosystem, not everyone is motivated by product costs and some consumers cannot opt-out of using speech technology. This implies that people might use and purchase AI-based technology regardless of their trust towards it, making it very challenging to measure or model trust.

### *2.2. Legal protection and audio technologies*

Despite their ubiquitous use in personal and professional settings, audio technologies pose risks to privacy, data protection and individual freedoms (Dhiya'Mardhiyyah et al., 2023). When people use technology that captures their speech or audio such as AI voice assistants (Förster et al., 2023), speech-enabled dolls (Williams et al., 2018) and voice modification apps (Hawley and Hancock, 2024), their voice and speech can be potentially re-used for purposes that are not clearly identified, or misused in ways that cause harm as with surveillance and fraud (Urquhart et al., 2022). Speech data can reveal a person's identity through voice biometrics (Campbell, 1997) and also sensitive information as ethnic origin (accent) (Newman and Wu, 2011; Mohammad and Al-Ani, 2017) or health related information (Harel et al., 2004; An et al., 2018; Mitra and Shriberg, 2015). AI speech recognition systems often struggle to understand certain accents and dialects due to insufficient training data (Kim et al., 2024) leading to biased or inadequate products for consumers. Speech can reveal private and confidential information of the speaker and others in conversation or the background (Williams et al., 2023c, 2022), including the words and content (Williams et al., 2023b), the context, environment and interlocutors (Ntalampiras et al., 2012), and even bystanders who have not consented to their speech data processing (Welch and Williams, 2024). Nearly all of this information is considered to be *personal data* and is governed by data protection laws as with the General Data Protection Regulation (GDPR) (Nautsch et al., 2019).

Within GDPR, speech data can be qualified as biometric or special category data. The unauthorised capture and re-use of speech audio can infringe upon rights as well as privacy and reputation. Recently, speech and voice modification that are used in artistic projects and media (even using the voice of deceased persons from *digital remains* Stokes, 2015) has stirred a significant legal and ethical debate about the limits of creative audio modification with respect to a wide variety of individual rights (privacy, publicity, personality, copyright, intellectual property). The legal protection of these rights differs between different jurisdictions (e.g., U.S., U.K., and EU, etc.), which can make it challenging for companies to comply with regulations that protect consumer needs.

The impact of continuous AI innovation has added even more socio-legal challenges. For example, voice actors face job displacement due to the synthetic use of their voices in the creative industries and content creation as advertisements or as social media content (Hutiri et al., 2024). Voice recognition systems have shown to demonstrate bias for speakers of different speaking styles, accents and races (Tatman, 2017; Koenecke et al., 2020). Voice recognition systems are often used for various processes such as job recruitment, immigration or residence applications and travel. With the potential for errors in accuracy, automated decision making could have serious consequences for people's lives and violate their right to non discrimination (Bajorek, 2019; Zuiderveen Borgesius, 2018).

### 2.3. Speech enhancement (A.1, F.1)

Speech from a desired speaker is often recorded in suboptimal acoustic environments, with the presence of background noise, speech from other people, and reverberation. Speech enhancement (SE) is the process of improving the quality and intelligibility of degraded speech signals, which may include removing undesired background noise and isolating the speech signal (Nuthakki et al., 2022). Such processing is useful for hearing aids (Williams et al., 2023a), to improve the accuracy of automatic speech recognition applications, and to ease the cognitive burden on listeners (Schröter et al., 2022). Speech enhancement systems are widely deployed in commercial products, including phone calls, VoIP (Microsoft Teams, Zoom, FaceTime, etc.), video conferencing, and hearing aids. As such, many of these SE systems are concerned with low latency real-time processing (in some cases, low-complexity implementations on embedded devices) while preserving the naturalness and improving the intelligibility of speech. In addition to Audio-Only Speech Enhancement (AOSE), Audio–Visual Speech Enhancement (AVSE) also gained popularity in recent years. AVSE leverages visual (facial/lip movement) information to enhance target speakers while rejecting off-camera interference speakers. Examples of deployed commercial AVSE systems include the on-screen target speaker enhancement of YouTube Shorts videos and AVSE for video recordings on smart phones (e.g., iPhone 16). In general, AOSE or AVSE systems are deployed in commercial products without explicitly notifying users. AOSE approaches range from classical spectral subtraction (removing an estimate of distortion components in the received signal) (Yan et al., 2020) to recent deep neural network (DNN) methods, which have minimal assumptions about the type of noise. Most state-of-the-art AOSE models use the short-time Fourier transform (STFT) representation and estimate a Time–Frequency (TF) mask using a deep neural network, either real-valued masks (Romero and Speckbacher, 2024) or complex masks (Zhang et al., 2022). Recently, researchers also investigated studio voice, a related task that aims to improve consumer-grade recordings (which suffer from moderate noise, reverb, and EQ distortion) to professional studio quality using speech re-synthesis based (generative) approaches (Su et al., 2021; Babaev et al., 2024). AVSE architectures are similar to AOSE except it uses a DNN based visual encoder to process the information from facial/lip movements and learns audio–visual correspondence.

### 2.4. Automatic speech recognition (A.1)

Automatic speech recognition (ASR) aims to transcribe speech sound captured in audio form into written text as words or phones. Speech recognition is challenging because of the inherent diversity in speech signals ranging from pronunciation and accent to natural disfluencies like pauses and repairs. Currently, ASR is used widely across many applications, such as aggregating weather reports, automatic call centre handling, and voice assistants (Alharbi et al., 2021). Recent notable advances in ASR include “Listen, Attend, and Spell” (LAS) (Chan et al., 2016), connectionist temporal classification (CTC) (Graves et al., 2006), recurrent neural networks with transducers (RNN-T) (Graves et al., 2013), conformer models (Gulati et al., 2020), weakly-supervised (Radford et al., 2023), and self-supervised (Baevski et al., 2020) approaches.

### 2.5. Video understanding (A.2)

Video understanding encompasses a variety of tasks that enable a human-like semantic understanding of video content (Lin et al., 2019), including video classification (Gupta et al., 2023), object detection and tracking (Nandhini and Thinakaran, 2023), video segmentation (Cheng et al., 2023), video summarisation (Saini et al., 2023), video captioning (Yang et al., 2023), and audio–visual question answering (Yang et al., 2022). Video and image understanding is increasingly becoming multimodal as the technology takes advantage of visual information in conjunction with text and audio information, especially in the scenarios where ego-centric video and audio capture is employed to assist someone in their interactions with the world in real-time. One example product vision is Google’s Project Astra (Google, 2025), “A research prototype exploring future capabilities of a universal AI assistant”. Users can wear a pair of eye glasses equipped with camera(s) and microphone(s) to interact with the world through a multimodal AI assistant, which has the ability to understand visual, audio, and language information. It also has memory capabilities by recording multimodal information and storing this data to enable easy information retrieval and to facilitate the capabilities of large language models (LLM) and AI agents.

### 2.6. Audio scene description (A.2)

Audio scene description is an established research area that is explored through communities with shared challenges like Detection and Classification of Acoustic Scenes and Events (DCASE) (DCASE, 2025) in the past decade (Khandelwal et al., 2024). DCASE consists of a variety of tasks for automatically detecting, tracking, and interpreting soundscapes and includes elements of acoustic scene classification, sound event detection, and automatic audio captioning. The goal of the latter is to generate a natural language sentence that describes the events in an audio recording, using deep neural networks. Audio scene description technology can be a combination of multimodal technologies of audio captioning, video understanding, and natural language generation.

### 2.7. Text-to-speech synthesis (A.3, B.1)

On the opposite end of the spectrum to ASR, text-to-speech (TTS) synthesis generates speech sounds and words from text input (Hunt and Black, 1996). TTS synthesis often incorporates additional information about grammar (Delmonte et al., 1986), prosody (Roekhaut et al., 2010; Hirschberg, 2006), and target speaker features (Fan et al., 2015). TTS is a classic building block of automatic dialogue systems. In recent years, TTS has flourished due to advances in vocoders such as WaveNet (van den Oord et al., 2016), WaveRNN (Kalchbrenner et al., 2018), HiFiGAN (Kong et al., 2020), and diffusion models (Chen et al., 2022). Further, there have been significant advances for generating mel spectrograms such as with Tacotron (Wang et al., 2017; Weiss et al., 2021). These technologies now include some limited ability to control how a speaker sounds in terms of prosody and emotion, though this remains a challenging research area. Many LLM products today such as ChatGPT feature a voice interface with natural-sounding synthetic speech. Developers can choose from a wide range of commercial providers that have state-of-the-art TTS APIs, such as Google TTS, IBM, Amazon Polly, ElevenLabs, and most recently, OpenAI TTS.

### 2.8. Voice cloning (A.3, C.1, F.2)

Advancements in neural vocoders have made voice cloning easier, and often with little data from the target speaker. Given a few seconds to a few minutes of target speaker voice embeddings (Lian et al., 2022; Dang et al., 2022), a trained neural network is able to generate realistic speech (or singing (Wang et al., 2023)) for any speaker. Voice cloning, like speaker control in TTS synthesis, relies on speech disentanglement wherein the content and speaker identity are separated (Williams et al., 2021).

### 2.9. Task-based agents and natural language understanding (D.1)

Modern voice assistants are virtual conversational agents designed to perform specific tasks based on user input. Such examples include Amazon Alexa, Apple Siri, and Google Assistant. These virtual personal assistants (VPAs) carry out routine tasks such as playing music, weather forecast, smart home management, traffic updates, and updating appointment calendars. The backend engine that powers these commercial VPAs include modular pipeline components such as ASR, NLU (Natural Language Understanding), dialogue state tracking and management, text-based NLG (Natural Language Generation), etc. Speech synthesis allows VPAs to carry on a conversation with users. However, VPAs are distinct from recent LLM-based chatbots such as ChatGPT (OpenAI, 2025) in that they are designed first and foremost to carry out real-world tasks.

### 2.10. Machine translation (E.1)

Machine translation (MT) automatically translates text from a source language (e.g., English) to a target language (e.g., Spanish). In the last decade, interest and research in neural machine translation has driven the advancement of not only higher quality MT engines but also contributed to advances in deep learning. One such advancement is the transformer architecture for MT (Vaswani et al., 2017) which currently powers almost of all of the current LLM architectures.

### 2.11. Speech-to-speech translation (E.1, F.1)

Speech-to-speech simultaneous translation allows people to communicate directly with each other using their respective native or preferred language. The system takes speech in a source language as input and then outputs speech in a desired target language (Williams et al., 2013). A traditional pipeline approach uses separate modules beginning with ASR to transcribe speech as text in the source language, then applies statistical MT alignments to translate text. Optionally, the translated text can be re-synthesised using TTS. However, this approach is known to be error prone. Instead, speech-to-speech simultaneous translation system translates phones from speech input into phones as speech output using neural networks and does not require intermediary text processing (Ren et al., 2020).

## 3. Methodology

In order to understand the public's preferences and identify the key requirements needed to build trust in future speech and audio technologies, we conducted a large-scale survey probing public perceptions of two types of socio-technical scenarios across a variety of application domains: (1) existing speech technology systems (Table 1), and (2) emerging or hypothetical speech technology capabilities (Table 2) that may help the public envisage issues in the wake of fast-paced speech technology developments. By enabling people to respond to speculative and existing use cases, we can begin to understand the behavioural attributes of this ecosystem (social + technical), leading to a greater appreciation of related issues and concerns, and ultimately to the development of enhanced trust and responsible speech technologies. In this section on methodology, we discuss how we developed the scenarios used in our survey (Section 3.1), how the survey was administered (Section 3.2), a statement on data availability and ethics (Section 3.3), a description of the demographic data that was collected (Section 3.4), our technique for translating response scales (Section 3.5), our process for developing the survey variables (Section 3.6), and we introduce four estimated variables that we utilised for addressing our research questions (Section 3.7–3.10). Our four estimated survey variables cover awareness of rights (Section 3.7), perceived responsibility (Section 3.8), exposure to risks (Section 3.9), and perceived trust (Section 3.10). The survey questions that entail each of the four variables are provided verbatim in Appendix B.

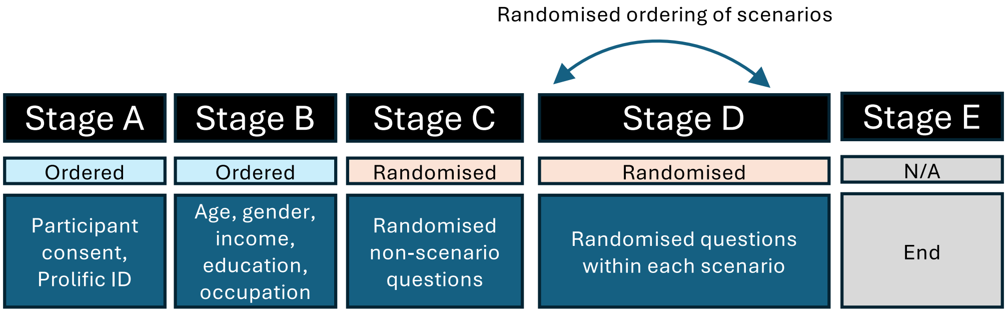

Randomised ordering of scenarios
Stage A
Ordered
Participant consent, Prolific ID
Stage B
Ordered
Age, gender, income, education, occupation
Stage C
Randomised
Randomised non-scenario questions
Stage D
Randomised
Randomised questions within each scenario
Stage E
N/A
End

**Fig. 1.** Overview of the survey flow in 5 stages, indicating which stages presented the question items in order versus randomised. The scenarios in Stage D included two types of randomisation: the presentation of scenarios, and the questions within a given scenario.

### *3.1. Developing socio-technical scenarios*

We utilised 'Responsible Research and Innovation (RRI) Prompts and Practice Cards' (Portillo et al., 2023) that allowed us to brainstorm our potential socio-technical scenarios in the areas of audio and speech technology using the Miro collaborative visualisation tool (MIRO, 2025). During this brainstorming process, the authors – a mix of industry and academic professionals – brainstormed and refined priority research areas and cross-disciplinary impact. The scenarios were later grouped into thematic application categories, which is how we arrived at our set of scenarios used in this work. Sample scenarios A–F are provided in Table 1 (see also Appendix A, Table A.5), and sample scenarios G–I are provided in Table 2 (see also Appendix A, Table A.6). The RRI discussion prompt cards are based on the AREA (Anticipate, Reflect, Engage, Act) framework for RRI (Owen et al., 2013), where each card emphasises one element of the AREA framework focused on "Purpose, Product, People, and Process", known among RRI experts as the Four Ps (Ameen et al., 2022).

### *3.2. Methodology for survey*

We hosted the survey using Qualtrics and recruited participants through the Prolific online platform. Please refer to Section 3.3 for information about participant consent, ethics approval, and data availability. A workflow of the survey is provided in Fig. 1 based on five stages. In Stage A, the participants signed up to the survey, read the Participant Information Sheet, provided their consent, and entered their Prolific identification number. In Stage B, the participants self-reported their demographical information for age, gender, income level, education level, and occupation. In Stage C, participants were asked questions (not directly linked to scenarios) such as how often they use speech technologies, whether they read terms and conditions, and if they are familiar with laws and regulations. In Stage D, the presentation order of scenarios was randomised, and within-scenario questions were also randomised. In Stage E, participants reached the end of the survey and were prompted to use a link to connect their Qualtrics survey completion with the Prolific platform for payment.

All survey participants were registered on the Prolific platform as being at least 18 years of age and residents of the UK. All of the survey data from the participants was anonymised through the Prolific platform and we did not collect any personally identifiable information. The survey consisted of 242 screens including the scenario descriptions and question items. The average (median) completion time for the survey was 46 min 17 s. Our judgement was that it would be suspicious to thoughtfully advance through all 242 screens in the Qualtrics survey within 10 min, as that would be an average of 2.4 s spent per screen, not allowing time to answer the question prompts even for a fast reader. Therefore respondents who completed in less than 10 min were not included in our final analysis, requiring us to collect more than 1,000 responses (initially 1,243) to meet our target. If we had included those who completed very quickly, such data points may have skewed our analysis to unrealistic and inaccurate results (Zhang and Conrad, 2014).

The choice of sample size is determined by taking into account the statistical validity and the practical feasibility of available resources and time. Large sample sizes are typically considered better because they reduce the likelihood of sampling errors and provide a more accurate representation of the population. However, this comes at the cost of more time, budget and the availability of the participants. We have therefore aimed for a sample size that provides a good balance between statistical significance and real world constraints of gathering more data.

Of the initially collected 1,243 responses, we omitted 10 who completed too quickly, 70 participants returned the survey to withdraw, and 15 accepted the survey but did not complete it (known as 'time-out' on the Prolific platform), resulting in an initial response rate of 92.3% (1,148/1,243). All 1,148 participants who completed the survey were paid a fixed reward of GBP 9.50. Since some people completed the survey faster than others, on average, this payment equates to GBP 8.98 per hour. We further eliminated 148 responses whose Prolific completion code could not be verified (known as 'no-code' on the Prolific platform), leaving us with a final response rate of 80.4% (1,000/1,243). Our remaining analysis hereafter considers these 1,000 participants. In this work, we focus our analysis on 46 question items that we carefully selected among the total 242 question items of the survey, to address our research questions (**RQ1–RQ4**). Of these 46 question items, 4 were related to demographics self-reporting. The remaining 42 question items are presented verbatim in Appendix B Tables B.7.

### 3.3. Data ethics and availability statement

In order to protect the rights of the study participants, the research questions and survey design were evaluated and approved by the Faculty Research Ethics Committee (FREC) at the University of Southampton.[2] Participants consented to disclosing their demographic details and to providing their survey responses, and could opt-out at any stage before submitting their completed survey. The participant responses were pseudonymised via a 24-character alphanumeric ID automatically provided by the Prolific recruitment platform at the time of collection and no personally identifiable information was collected. Participants were provided with contact information to submit questions or complaints to the researchers and the University of Southampton research ethics office. While our ethical process does not allow us to share original raw Qualtrics survey data, we have made available the report containing results for each survey question.[3]

### 3.4. Demographic information (DI)

The following demographic attributes of the participants were recorded based on self-reporting questions in our survey: age, highest level of education, occupation, annual household income, and gender. The questions came from Stage B of the survey workflow (Fig. 1). We did not collect information about race, religion, or other protected characteristics which are considered to be special category information under UK GDPR, and which would have required additional data governance and ethical approval. Visualisations of the distribution of demographic categories is provided in Appendix C.

### 3.5. Translation of question response scales

Consistency of measurement in our survey was achieved by translation of the user responses into a fixed 5-point Likert scale (Likert, 1932). We selected Likert over Likert-type (Boone Jr. and Boone, 2012) in order to develop the responses into a composite scale for generating quantitative measures of our estimated variables that will be presented in Section 3.7–3.10. Our main objective in developing the survey question items and response choices was to maintain semantic clarity for the participants. Therefore, some of our survey questions were well-suited as a direct yes/no/unsure response format, which places the responses into a 2-point and 3-point scale. Other questions were semantically justified by a 5-point scale (e.g., strongly negative, weakly negative, neutral, weakly positive, strongly positive). For consistency, all question items where the response format required a scale of $k < 5$ underwent linear-based translation from a $k$-point scale to a 5-point scale, where ordered numerical values in the range of 1 (*strong negative*) to 5 (*strong positive*) were attached to the response categories. The $k$ in a $k$-point scale refers to the number of response options provided, which can be any integer greater than or equal to 2. Commonly used values for k include 3, 5, and 7. For semantic clarity to survey respondents, we had initially used 2, 3 and 5 response options for the questions posed in the survey. These responses were later mapped onto a 5-point scale. For purposes of reproducibility, we report the ordered numerical values for each survey question in Appendix B Tables B.7. There are known trade-offs for using linear-based translation of Likert scales. However as we do not attempt to characterise individual responses, we argue that the problem of tied scores is a non-issue in our survey, and that instead our translation has the simple effect of weighting semantically strong questions (Chakrabartty, 2020). We conduct a preliminary statistical analysis of the suitability of our variables later in Section 4.1.

### 3.6. Development of survey variables

The four new variables in our survey will be outlined in more detail in the following Section 3.7–3.10. Here we describe the process by which the variables were developed. We followed a systematic procedure with five main steps (Aithal and Aithal, 2020; Converse and Presser, 1986) in sequence: (1) clear understanding of the survey aim and research questions of interest, (2) designing specific sociotechnical scenarios that needed to be explored to address the research questions, (3) translating these scenarios into measurable variables, while classifying them as either dependent (the outcome being measured) or independent (factors that might influence the outcome), (4) selection of an appropriate type of scale to semantically measure each type of variable (nominal, ordinal, etc.), (5) development of survey questions that could operationalise the variables ensuring their clarity, relevance and accuracy. For each variable that we introduced, we calculated the internal consistency-reliability of the questions using McDonald's Omega ($\Omega$) score (Stensen and Lydersen, 2022). An $\Omega$ above 0.70 indicates that our scale for each variable is measuring what it intends to measure, and that there are a sufficient number of well-correlated questions to measure the underlying concept of the latent variable.

[2] Reference FREC Approval: ERGO #78339.

[3] https://doi.org/10.5281/zenodo.16919218

### 3.7. Perceived awareness of rights (PAoR)

We developed the variable *perceived awareness of rights* (PAoR) to estimate the public's awareness of their rights and obligations regarding the use of speech technologies, which aligns with **RQ1**. The 13 question items posed in our survey that relate to rights (see Appendix B, Table B.7) encapsulated user opinions on privacy, data protection rights, copyright, publicity issues and consumer rights. The question items used for this variable came from Stage C and Stage D. All 13 items were mapped into on a 5-point Likert scale. The total scale score was obtained by taking the average across all of the 13 question items. The internal consistency score across the 13 items is $\Omega = 0.72$, which demonstrates that the items on a 5-point Likert scale are correlated without redundancy, and are measuring the same latent variable from here referred to as PAoR.

### 3.8. Perceived responsibility (PR)

In order to estimate participant understanding of their responsibility in using speech technology systems, we introduce the variable of *perceived responsibility* (PR). Responsibility in this context relates to compliance for rules and regulations, taking preventive and precautionary behaviours to protect one's own and other peoples' rights. We identified 4 question items in our survey (see Appendix B, Table B.8) related to responsibility for addressing **RQ2**. The question items used for this variable came from Stage C only, and responses were mapped into a 5-point Likert scale. The internal reliability of these 4 items on this scale is $\Omega = 0.80$, reflecting that the items on a 5-point Likert scale are correlated without redundancy, and are measuring the same latent variable from here referred to as PR.

### 3.9. Exposure to risks (EtR)

We also asked participants if they felt susceptible to harms such as violations of their privacy, compromise of their voice data security, misuse of their voice data for unauthorised biometric surveillance, etc., based on their experiences of using speech technology products. To gauge this, we introduce a variable called *exposure to risks* (EtR) which estimates participant's perceived vulnerability. The 13 question items about vulnerabilities discussed in our survey (see Appendix B, ) are related to data security, protection, and privacy and align to **RQ3**. The question items used for this variable came from Stage D only and responses were mapped into a 5-point Likert scale. The internal consistency of these 13 items is $\Omega = 0.73$, reflecting that the items on a 5-point Likert scale are correlated without redundancy, and are measuring the same latent variable from here referred to as EtR.

### 3.10. Perceived trust (PT)

We estimated participant's *perceived trust* (PT) by presenting a variety of questions that probed them to consider how they feel about trusting speech technologies, addressing **RQ2** and **RQ3**. The scenarios were addressed with 12 question items (see Appendix B, ) from Stage D only and responses were mapped into a 5-point Likert scale. The internal reliability for these 12 items is $\Omega = 0.75$, indicating that the items on a 5-point Likert scale are correlated without redundancy, and are measuring the same latent variable from here referred to as PT.

## 4. Data analysis and results

Our analysis begins with an assessment of the statistical suitability of the four variables PAoR, PR, EtR, and PT (Section 4.1). We then examine their correlations (Section 4.2), explore the impact of demographic factors (Section 4.3), and the correlation between our estimated variables (Section 4.4). All of our statistical analysis was conducted using SPSS (v.29) (IBM Corp, 2022).

### 4.1. Preliminary assessment of variables

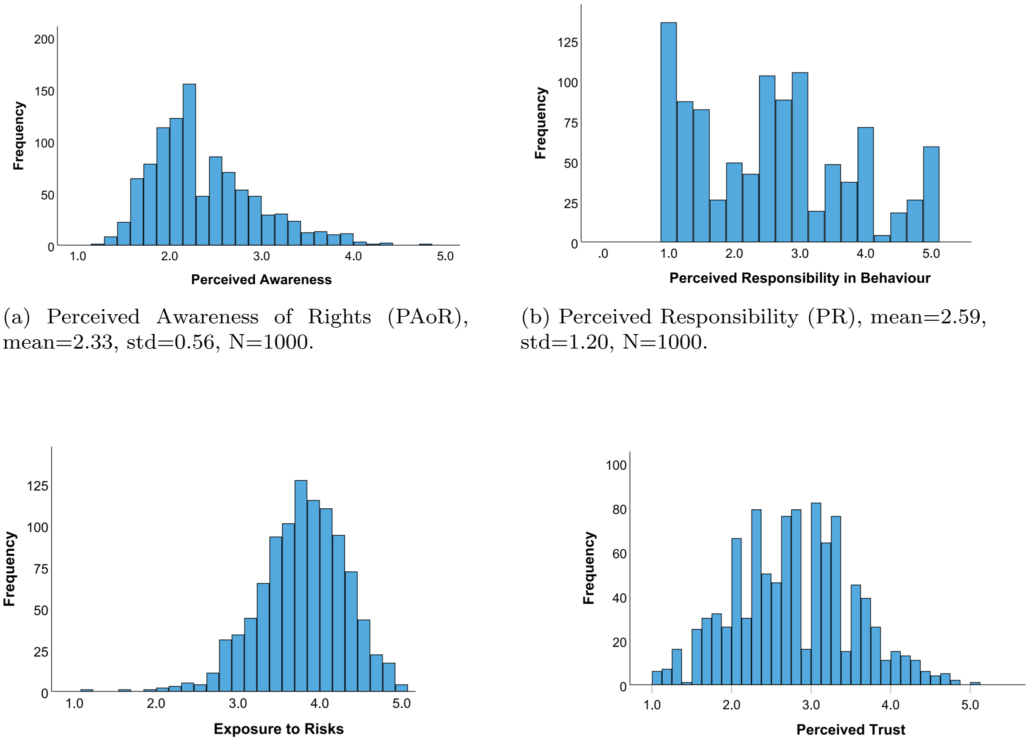
Building upon our earlier discussion of how the question response scales were translated into a 5-point Likert scale (Section 3.5), we further calculated descriptive statistics of the variable range, minimum, maximum, mean, standard deviation. We then examined the skewness and kurtosis (Hair Jr. et al., 2013) for each variable by comparing each distribution with the normal distribution. Skewness measures the asymmetry of a data distribution and it helps us to understand which side (left or right) the tail of the distribution is longer or wider. The skewness for a normal distribution is zero, therefore if our variables are also symmetric then they should exhibit a skewness value close to zero. Kurtosis indicates how peaked or flat a distribution is relative to the normal distribution. From our estimated variables, all four have a skewness and kurtosis between $-1$ and $+1$. This range indicates that the data is approximately normal, or 'close enough' for additional statistical tests to work properly, such as ANOVA and t-test (Glass et al., 1972; Harwell, 1992; Lix et al., 1996). We summarise the statistics for our estimated variables in Table 3. The skewness, kurtosis, and $\Omega$ scores indicate that our decision to linearly translate the scales of our $k$-point question responses into a 5-point scale did not have an obvious negative impact on our ability to conduct further analysis.

In order to visually inspect the normality of the probability distribution for each variable, we have plotted histograms that depict the frequency distribution of the observed variables. Visualisation through histograms can highlight any gaps in the data or the existence of outliers in the distribution tails (Ghasemi and Zahediasl, 2012). From the histograms shown in Fig. 2, one cannot notice any significant departure from normality.

**Table 3**

Statistical summary for the four estimated variables (PAoR, PR, EtR, and PT) after mapping responses into a 5-point Likert scale: number of responses (N), number of question items, range of values for responses, mean, standard deviation, internal consistency ($\Omega$), skewness statistic, and kurtosis.

| | N | #Q | Range | Mean | Std. | $\Omega$ | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|
| **PAoR** | 1000 | 13 | 3.5 | 2.33 | 0.56 | 0.72 | 0.84 | 0.56 |
| **PR** | 1000 | 4 | 4 | 2.58 | 1.19 | 0.80 | 0.37 | -0.82 |
| **EtR** | 1000 | 13 | 3.84 | 3.75 | 0.52 | 0.73 | -0.50 | 0.81 |
| **PT** | 1000 | 12 | 4 5 | 2.77 | 0.74 | 0.75 | 0.12 | -0.31 |



(a) Perceived Awareness of Rights (PAoR), mean=2.33, std=0.56, N=1000.

(b) Perceived Responsibility (PR), mean=2.59, std=1.20, N=1000.

(c) Exposure to Risks (EtR), mean=3.75, std=0.52, N=1000.

(d) Perceived Trust (PT), mean=2.77, std=0.74, N=1000.

**Fig. 2.** Histogram distributions for each variable after mapping to 5-point Likert scale.

The distribution of variable EtR (Fig. 2(c)) has a long left tail which was also captured by the negative skewness value. The remaining three variables (Fig. 2(a), 2(b), and 2(d)) have positive skewness and therefore exhibit a longer right tail, comparatively. The distribution of the variables also provides intuition about general tendency. For example, the majority of our participants declared that they are not sufficiently aware of their rights (Fig. 2(a)). Very few people claimed high levels of trust in speech technology (Fig. 2(d)). Most people have reflected a less responsible attitude towards the use of speech data (Fig. 2(b)) and most of them believe that their exposure to risks is high (Fig. 2(c)).

### *4.2. Correlation between variables*

In order to identify associations between our four estimated variables, we computed Pearson correlation coefficient ($\rho$) as shown in Table 4, with significance at p = 0.01 (two tailed). The values for $\rho$ inform the direction of correlation (negative vs. positive), and strength of the linear relationship (strong vs. weak) between each of the variables. From the values shown in Table 4, we enumerate our findings with commentary:

- **Finding #1:** *Perceived responsibility* (PR) has a moderately positive correlation ($\rho$=0.46) to *perceived awareness of rights* (PAoR). **Response to RQ1:** Although the responsible use of speech technology seems positively correlated to the awareness of legal and human rights, it does not imply a causal relationship. The majority of our participants have showcased insufficient knowledge of legal and human rights but have also shown high interest in knowing more about these. Participants have also expressed the importance of having regulations in different socio-technical contexts. Given the positive correlation, members of the public

**Table 4**
Pearson correlation coefficient ($\rho$) between each of the four estimated variables with significance at p = 0.01 (two-tailed).

| | PR | PAoR | PT | EtR |
|---|---|---|---|---|
| **PR** | 1 | | | |
| **PAoR** | 0.46 | 1 | | |
| **PT** | -0.05 | 0.13 | 1 | |
| **EtR** | 0.13 | -0.02 | -0.49 | 1 |

may benefit from proactive engagement through awareness campaigns to increase collective awareness of human and legal rights and may benefit from a deeper understanding of the consequences for violating such rights.

- **Finding #2:** *Perceived trust* (PT) and *perceived awareness of rights* (PAoR) are weakly positively correlated ($\rho$=0.13). **Response to RQ2:** Further exploration through causal inference techniques can suggest if the knowledge of legal and human rights has the ability to influence how people trust speech technology. Understanding what enables people to trust speech technology could have a positive economic impact, as well as support continued speech technology innovation across many sectors.
- **Finding #3:** *Perceived trust* (PT) has a moderately negative correlation ($\rho$=−0.49) with *exposure to risks* (EtR). **Response to RQ3:** Due to the risks and vulnerabilities associated with voice data, companies often attempt implement data privacy policies, security measures for risk mitigation, routine data governance, and other activities to achieve compliance with GDPR. However, such efforts may not necessarily win consumer trust. There may be other factors to uncover in order to explain why consumer understanding of risks leads to a loss of trust.
- **Finding #4:** *Exposure to risks* (EtR) has a weak negative correlation ($\rho$=−0.02) to *perceived awareness of rights* (PAoR). **Comment:** If this relationship of negative correlation is proven to be statistically significant (explored later in Section 4.4), it would imply without causality that digital harm tends to occur where there is less awareness of legal rights.
- **Finding #5:** *Perceived responsibility* (PR) has a weak negative correlation ($\rho$=−0.05) to *perceived trust* (PT). **Comment:** If found to be statistically significant, this correlation would imply without causation that people who behave more responsibly around speech technology also tend to trust it less.
- **Finding #6:** *Perceived responsibility* (PR) has a weak positive correlation ($\rho$=0.13) to *exposure to risks* (EtR). **Comment:** If found to be statistically significant, this correlation implies, without causation, that individuals who have some understanding or exposure to the risks associated with speech technologies and voice data, may also act more responsibly across socio-technical scenarios. Building from the context of **Finding #3** and **RQ3**, it makes sense that participants would have less trust for speech technology if the onus falls upon consumers rather than industry developers or governments.

### 4.3. Impact of demographic factors

To address **RQ4** we used one-way ANOVA reporting F-statistic evaluated at significance $\alpha = 0.001$ to determine if any impact of demographic factors was systematic or by chance. To perform this statistical analysis, each of the four estimated variables (PAoR, PR, EtR, and PT) were treated as dependent variables and demographic attributes (age, gender, education, and occupation) were treated as independent variables. The categorical nature of the demographic attributes also allows us to further compare the sub-groups within each type of demographic attribute. The one-way ANOVA test uses the F-statistic for statistical significance. A larger F-value means that it is more likely that the variation associated with the independent variable is real and not due to chance. If the variation for demographic attributes is not due to chance, we can draw conclusions about the interplay between demographics and our four estimated variables for awareness, responsibility, risks, and trust. Such information is valuable to principles of AI safety as well as technology developers who may want to occupy all corners of the consumer market.

#### 4.3.1. Influence of gender

The five categories for self-reported gender from our survey included: male, female, non-binary/non-conforming, transgender, and undisclosed. We found that gender has a statistically significant impact on PAoR ($F$= 8.09, $p < 0.001$), as well as on PT ($F$= 8.32, $p < 0.001$). However, we were not able to identify exactly which genders amongst our categories created this impact via Tukey post-hoc tests because at least one category (transgender) had fewer than two instances. We found no statistically significant impact for gender on PR ($F$= 1.42, $p = 0.22 > 0.001$), or on EtR ($F$= 1.89, $p = 0.10 > 0.001$). Therefore any observable influence of gender on PR or EtR is purely by chance. Further detailed summaries of the ANOVA F-statistics are provided in Appendix D for PAoR in Table D.11 and PT in Table D.12 as these are the variables which were found to be statistically significant.

#### 4.3.2. Influence of education

Based on seven self-reported educational categories from within the UK system (GCSE, A-levels, trade training, bachelors degree, masters degree, PhD, and undisclosed), we found that education levels had no statistically significant impact on any of our estimated variables: PAoR ($p = 0.05 > 0.001$), PR ($p = 0.06 > 0.001$), EtR ($p = 0.51 > 0.001$), or PT ($p = 0.64 > 0.001$). We provide details of the ANOVA effect sizes and 95% confidence intervals for education with each survey variable in Table D.17a and D.17b.

### 4.3.3. *Influence of age*

Based on eight self-reported age categories of the participants (18–25, 26–35, 36–45, 46–55, 56–65, 66–75, 76+, undisclosed), we found that age had no statistically significant impact on any of our estimated variables: PAoR ($p = 0.42 > 0.001$), PR ($p = 0.94 > 0.001$), EtR ($p = 0.15 > 0.001$), or PT ($p = 0.37 > 0.001$). We provide details of the ANOVA effect size and 95% confidence intervals for age with each variable in Table D.18a and D.18b.

### 4.4. *Correlation between estimated variables*

Given our initial findings from Pearson's correlation coefficient ($\rho$) earlier in Section 4.2, we further investigated the statistical significance of these findings. Even for $\rho$ values that are considered weak, as was the case for most of the correlations between our estimated variables, statistical significance can inform whether the probability of observing those correlations is high or low. In general, larger values for $N$ survey responses facilitate the ability to identify significance, even for weak correlations. We identified statistically significant relationships (p = 0.01, two-tailed) between PR and PAoR (**Finding #1**) PAoR and PT (**Finding #2**), between PT and EtR (**Finding #3**), and between PR and EtR (**Finding #6**). We provide detailed summaries for variables which were found to be statistically significant in Appendix D, Tables D.13–D.16. In these detailed statistical summaries, we report the Pearson correlation, sum of squares, and covariance. The Pearson correlation was explained earlier in Section 4.2. The sum of squares indicates variability present in the data. And the covariance indicates how variables tend to change together. A positive covariance indicates that if one variable is above its mean, then the other variable is above its mean as well. A negative covariance value indicates that the variables do not change in a corresponding way. Positive covariance is observed between PAoR and PT, between PR and EtR, and between PAoR and PR. This suggests that our variables have captured relationships between responsibility and trust. However, negative covariance was observed between PT and EtR which suggests increased risk exposure corresponds to lower trust. We did not find statistical significance between PAoR and EtR (**Finding #4**), or between PT and PR (**Finding #5**).

## 5. Discussion of findings

From our survey design, methodology, and analysis, we had several findings that were validated by statistical significance tests. Our findings enable us to revisit the implications of our research questions, as well as to propose ideas for a diverse readership regarding how we may go forward with speech technology innovation in a manner that is safe and beneficial. We summarise our main findings in this section, mapped into each of our original research questions **RQ1–RQ4**.

### 5.1. *Findings for responsible use*

**RQ1**: Is appropriate and responsible *use* of speech technology influenced by the public's understanding of legal rights and human rights?

**Discussion**: There is a relationship between how people understand their rights and whether or not they use speech technology responsibly. While this correlation does not imply causality, such a relationship could potentially be leveraged to try and increase adoption of responsible practice when using speech technology with the aim to reduce societal harms, including harms caused by deepfakes. Responsible use, as reflected in our scenarios, involves more than just the user (or survey respondent) because it may involve bystanders or entire groups of users. The implication is that responsible use could be influenced by improving the public's awareness of their rights.

### 5.2. *Findings for trust*

**RQ2**: Is *trust* in speech technology affected by the public's understanding of legal rights and human rights?

**Discussion**: While our results do not imply causality, we found that trust and rights are strongly correlated. Improving upon the public's awareness of their rights, alongside potentially protective regulation or legislation, could play a role in shaping public trust towards speech technology. The general consequences of strong public trust are far-reaching and could contribute towards the UK's ambitious AI ecosystem goals at the national and global levels. Strong public trust also promotes AI technology uptake, which in turn helps to drive innovation. Awareness of rights is especially important in matters arising from opt-in cultures at professional organisations, where employees may feel pressured or be required to adopt speech technologies as a condition of their employment.

### 5.3. *Findings for risks*

**RQ3**: Are there any correlations between trust in speech technology and the public's understanding of potential risk exposures?

**Discussion**: Excessive vulnerability or exposure to voice security and voice privacy harms are correlated with levels of trust in speech technology. Individuals who experience fewer harms might be more likely to trust (as well as purchase, use, and share) a variety of future speech technology applications. While not identified as a causal relationship, public awareness of voice security and privacy risks are correlated with more responsible behaviours, and vice versa. This finding motivates the need for increasing public awareness of risks, especially to allow consumers to make informed decisions about the types of speech technologies that they engage with. While this may seem counter-intuitive for industry, improved transparency surrounding risks may have a positive impact on product uptake and brand reputation.

### 5.4. Findings for demographics

**RQ4**: Do perceptions of exposure to hazards, knowledge of human rights and the law, trust in speech and voice technology, and responsible behaviour depend on individual's demographic attributes such as age, gender and level of education?

**Discussion**: Among the demographic variables that we captured in our survey, gender had a statistically significant role for awareness of rights (PAoR) and trust (PT). We were not able to complete a deeper analysis with statistical significance to identify which gender(s) are most correlated to our variables due to the distributions in our sample population. Since we did find gender to be correlated, we encourage future work to look closer in this direction, especially as a means to inform potential innovation streams or market gaps. Factors such as age and education were not correlated our other variables. The lack of correlation for age and education level might indicate that issues of trust, risks, responsibility and rights are widespread across generations and socioeconomic strata. There are emerging studies exploring how different age groups benefit from or interact with AI (Lüdemann et al., 2024). Fewer studies explore age as a condition of algorithmic bias and fairness. Participants over the age of 56 made up nearly 20% of our survey responses. Our finding that age is not a significant demographic factor for trust towards speech technology is contrary to commentary from recent work (Stypinska, 2023) that speculated older adults would likely experience ageism when interacting with conversational chatbots. It is possible that additional statistical analysis beyond ANOVA may reveal correlations, especially after increasing the sample size of 1,000 participants by several orders of magnitude (Glass et al., 1972; Harwell, 1992; Lix et al., 1996). As noted in prior studies (Stypinska, 2023), we agree that future work should investigate the impacts of speech technology on older adults in greater depth. The majority of survey participants indicated that they are not sufficiently aware of their rights. As our survey has shown, awareness of rights is strongly correlated with trust and responsible behaviour for speech technology. Therefore, we encourage policymakers and regulators alike to increase their engagements for public diplomacy and to have a pro-active role in fostering public trust.

## 6. Discussion of limitations

We have identified several limitations of this work and address them categorically in terms of geographical region, technology, and survey methodology. First, this study was limited to the United Kingdom. It is reasonable to expect that access and exposure to speech technology in personal and professional contexts may differ between the United Kingdom and other regions of the world. We suggest that this work be replicated to study other regions and cultures, especially to compare across cultures. Second, we studied two broad categories of speech technology scenarios: real existing technology and imagined technology of the future. We are unable to judge through a survey the depth of understanding that the public has about how these technologies work. Therefore our findings are a reflection of how people perceive the technology, which may differ greatly from person to person. A more targeted pool of respondents, an interactive survey, or a focus group may reveal further differences among respondents based on their depth of technical understanding. While we did not find significant correlations for education level or age, it remains to be explored in greater depth whether socioeconomic status has a role in trust. A recent assessment of the impact of socioeconomic status and generative AI, for example, foreshadows that lower socioeconomic groups may experience job displacement (Capraro et al., 2024) which could lead to loss of trust. Finally, we comment on our survey methodology. Our survey could not include all of the speech technologies that currently exist, or the vast variety of scenarios for which they may be used. Many speech technologies are still under development by academics and researchers, and not yet known to the general public. Alongside a possible technical knowledge gap, adding more technologies or scenarios may be counter-productive due to survey length resulting in respondent fatigue or loss of interest (Yan et al., 2011). An approach could be developed to run several different disjoint surveys across a larger set of the population to address this limitation.

## 7. Conclusion

We have presented the first large-scale public survey with 1,000 participants to examine public attitudes in the United Kingdom for trust towards speech technology. We have demonstrated through a review of prior work and recent products, alongside contextualised scenarios, that the term *speech technology* has a broad meaning, even if it is significantly more narrow than *artificial intelligence*. We argue that a timely understanding of public trust towards speech technology feeds into the wider global conversation about AI safety, where nuances about perceived responsibility, perceived trust, awareness of rights, and exposures to risk and harm — may get lost in breadth of AI regulation. Or may be overpowered by a flood of new speech technology products for consumers.

Our survey design focused on existing commercial products, emerging capabilities, and hypothetical scenarios, based on the current state of speech technology research. Our development of these scenarios was rooted in the RRI AREA framework which allowed us to anticipate, reflect, engage, and act according to potential concerns that may impact how the public trusts different kinds of speech technology in their everyday life. In our methodology, we introduced four new variables to encapsulate aspects of public trust: *perceived awareness of rights* (PAoR), *perceived responsibility* (PR), *exposure to risks* (EtR), and *perceived trust* (PT). We used these variables to anchor our statistical analysis of survey responses to questions pertaining to each scenario. The survey data will be made available on request.

We explored four research questions to characterise the relationships between public perceptions of trust towards speech technology and fundamental human rights. Our survey design and analysis was tailored to answering these questions. We found that awareness of rights (PAoR) is a key factor that correlates to whether speech technology is used responsibly (PR), and whether it is trusted (PT). We also found that when people are exposed to risks or harms (EtR) this can affect their behaviour and use (PR) as well as their trust (PT). In our analysis of demographic factors, we identified that gender is a significant factor in awareness of rights (PAoR) and trust (PT) but not for exposure to risks and responsible use. We did not find significant effects for age or education level, but further surveys and statistical analysis testing may reveal effects in line with related works.

Our work contextualises speech technology and potential future regulation within the ongoing global discussions of AI regulation and AI safety. At the time of this writing, two prominent Global AI Safety Summits in the UK (UK Governmenet, 2025a) and South Korea (UK Government, 2025b) have already been held, alongside the newly enacted EU AI Act (Laux et al., 2024). However, not all global populations have equal opportunities to interact with speech technology in their personal or professional lives. This may be due to aspects of income inequality (many devices and subscriptions are expensive) or their intentional desire to avoid speech technology for everyday use. We propose future work that provides survey participants an opportunity to interact with a variety of speech technology capabilities and to repeat the survey questions in-context, and with a wider global population. For example, demonstrating to participants the current state of audio deepfakes by allowing them to clone their own voice in an interactive survey, before responding to the survey questions. Our work establishes a foundation for exploring the development of speech technology standards and regulations in greater depth, and our scenarios in context help to establish a framework for assessing public trust as speech technology continues to develop.

## CRediT authorship contribution statement

**Jennifer Williams:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Tayyaba Azim:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Anna-Maria Piskopani:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Richard Hyde:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Shuo Zhang:** Writing – review & editing, Writing – original draft, Conceptualization. **Zack Hodari:** Writing – review & editing, Writing – original draft, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jennifer Williams reports a relationship with The Alan Turing Institute that includes: consulting or advisory. Jennifer Williams reports a relationship with MyVoice AI that includes: employment. Jennifer Williams has patent #US20230186896A1 pending to MyVoice AI Ltd. Jennifer Williams has patent #US20220405363A1 pending to MyVoice AI Ltd. The corresponding author (Jennifer Williams) is a Co-Guest Editor for CSL Special Issue on Security and Privacy in Speech Communication. In regard to prior employment at MyVoice AI Ltd, Jennifer Williams was previously employed part-time (ending in Feb 2024) and does not have any remaining interactions nor any restrictive covenants, but there are two patents pending where Jennifer Williams is listed as an inventor and MyVoice AI Ltd is the assignee. Those two patents are topically related to one of the scenarios that was explored in the survey for the work in this manuscript, but otherwise the patents and IP are completely unrelated. Regarding consultancy, Jennifer Williams has recently been hired as a consultant to the Alan Turing Institute (July 2024) but has not yet undertaken any consultancy work at the time of this submission. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The authors acknowledge their funding source, the UK EPSRC Trustworthy Autonomous Systems Hub (EP/V00784X/1). We also express our sincere gratitude to Gopal Ramchurn, Sebastian Stein, Vahid Yazdanpanah, Chris Greenhalgh, Alan Chamberlain, Felipe Espic, Tom Iddon, and Shuvayanti Das, who advised us on the proposed work in the very early stages.

## Appendix A. Full scenario descriptions

See Tables A.5 and A.6.

## Appendix B. Survey prompts and responses

See Tables B.7 and B.8.

## Appendix C. Demographic information

See Fig. C.3 and Fig. C.4.

## Appendix D. Statistical analysis summaries and correlation between estimated variables

See Tables D.11–D.18.

**Table A.5**
Full scenarios for existing/known speech technologies.

| Application Domain | Existing Socio-Technical Scenarios | Relevance |
|---|---|---|
| A. Health Domain | Scenario A1 – Smart Hearing Aid: Jane is hard of hearing (almost completely deaf) and uses a special hearing aid to help her throughout the day. This hearing aid can enhance specific voices that are within 5–10 m from where she is standing. It can also generate transcripts of conversations, which she can access on her phone and read to herself | Speech Enhancement, Speech Recognition |
| | Scenario A2 – AI-Assistive Device for the Blind: Nancy is blind. She uses a special type of AI-assistive device that speaks out loud and describes her | Video Understanding, Audio Scene Description |
| | Scenario A3 – Text to Speech Interpreter: Hugh had a stroke and as a result now has aphasia, which has made him unable to speak. A device has reconstructed his voice from recordings and he uses it to converse with people in public, by typing in what he wants the device to say for him. | Speech Synthesis, Voice Cloning |
| B. Academia | Scenario B1 – Creating Audio Books: AI generative tools can also convert audio to text and text to audio. They could be used for artistic purposes such as creating audio books. | Speech Synthesis |
| C. Digital Voice Recording Market | Scenario C1 – Preserving Voice Memory of Loved Ones: They say that the first thing you forget about a person is the sound of their voice. Preserving people's voice, memories and personal stories can be valuable to families and next generations. Interviews with loved ones in combination with their favourite music can construct a voice and memories album. | Voice Cloning, Expressive speech synthesis |
| D. Smart Environments | Scenario D1 – Vehicle: While you are driving, you can use your smart vehicle's audio capture capabilities to create a grocery shopping list, order takeout, or buy an audio book from the car without taking your eyes off the road using a pre-installed voice assistant. Your voice assistant can also predict some of your needs and make suggestions or choose and shop for products using your prior history, lifestyle behaviours and patterns as well as your budget settings. | Task-based Agents, Natural Language Understanding |
| | Scenario D2 – Work: Your employer has decided to renovate the office or building that you work in and is planning to include smart AI technologies that are meant to facilitate productivity and ultimately reduce operational costs. They have installed microphones throughout the area. | |
| | Scenario D3 – Home: You have recently purchased a new voice assistant for the home. This is a device that can sit on the kitchen countertop or in the living room. You can speak naturally to it and ask questions, set reminders, play music, or get information from the internet. The voice assistant speaks back to you when you talk to it. | |
| E. Films Industry | Scenario E1 – Translation Services: Automatic translation services provide films and television shows that are not in your native language. For example, your preferred television news service or radio station is only provided in a rare foreign language, so you utilise a translation service. The translations happen automatically and instead of providing captions, the translated content is spoken just like a person is speaking. Still, you are told that an AI algorithm has performed the translation service. | Machine Translation, Speech-to-Speech Translation |
| F. Voice Conferencing | Scenario F1 – Work: The company you work for requires you to use a video conferencing tool to engage in a high-stakes meeting with an international business partner. | Speech enhancement, Speech-to-speech Translation |
| | Scenario F2 – Internet: You have posted a recording of your voice to the internet (Facebook, YouTube, TikTok, etc.). Someone else can use this recording to make new recordings which sound like you (i.e., using your voice). | Voice Cloning |

**Table A.6**
Full scenarios for imagined/hypothetical speech technologies.

| Application Domain | Hypothetical Socio-Technical Scenarios | Relevance |
|---|---|---|
| G. Interactive Children's Toys | Scenario G1 – AI Dolls: Someone gives your child an interactive doll as a present. It is an internet-connected doll equipped with speech recognition systems and AI based learning features, operating as an Internet of things (IoT) device. The doll remembers the child's play history and past conversations and can suggest new games and topics. The doll is named "Lucy" and she interacts with your children and can answer their questions. | Speech Recognition, Speech Synthesis, Task-based agents and NLU, Autonomous Robots |
| H. Health Domain | Scenario H1 – Health Monitoring App: You have recently been diagnosed with a breathing disorder that requires supervision by your medical doctor. Your doctor recommends a new AI-based app that uses audio from a smartphone or similar device, allowing instant feedback about the condition quickly and can alert if there is a medical emergency or not. | |
| | Scenario H2 – Social Interactive Robot for Nursing: Your elderly relative requires support throughout the day for reminders to take medicine, instructions for how to care for themselves, and social support to help with loneliness. Their doctor suggests using a new type of social-interactive robot to supplement the times when a home nurse is unavailable. This robot can speak and interact, and can even provide comforting support for loneliness or hold a casual conversation. The robot can also alert emergency services if there is a critical issue. The robot is always on and is recording audio in order to provide the necessary support and care that your relative needs. | Speech Recognition, Speech Synthesis, Task-based agents and NLU, Autonomous Robots |
| I. Audio Immersive Projects in Arts | Scenario I1 – Live Singing: An audio capture system is set up at a venue where you go to see a singer perform. It is an experimental performance where your voice is captured prior to the performance and then used to perform a series of songs. It's your voice and its sound that the singer uses to perform the first song - the singer sounds like you! | Voice Cloning |
| | Scenario I2 – Pets: A designer creates a system that can use your voice to talk to your pets while you're away to entertain them, comfort them and offer enrichment activities, such as talking or shouting commands while a robot arm throws a ball or plays with your pets favourite toy. | Speech Synthesis,Task-based agents and NLU, Audio Understanding |

**Table B.7**
Survey prompts for estimating perceived awareness of rights (PAoR).

| Survey Prompts | Responses |
|---|---|
| Stage C (no scenario):<br>1. Do you read the privacy policy in your products/services that you buy, or do you just agree? (Q137) | [Read (5), It depends (3), Don't Read (1)] |
| Stage C (no scenario):<br>2. Do you know if there are any laws or regulations that protect your right to privacy and data protection when using audio technologies? (Q138) | [Yes (5), Not sure (3), No(1)] |
| Stage C (no scenario):<br>3. Do you know your rights as a consumer when purchasing audio equipment, such as headphones or speakers? (Q151) | [Yes (5), No (1)] |
| Stage C (no scenario):<br>4. Are you aware of any organizations or resources that provide information or support regarding your rights as a consumer of audio technologies? (Q153) | [Yes (5), No (1)] |
| Stage C (no scenario):<br>5. Are you aware of any laws or regulations that govern the use of audio recordings as evidence in legal proceedings? (Q154) | [Yes (5), No (1)] |
| Stage C (no scenario):<br>6. Are you aware of anyone who has ever been denied their right to access their audio data due to their location or country? (Q155) | [Yes (5), No (1)] |

(*continued on next page*)

**Table B.7** (*continued*).

| Survey Prompts | Responses |
|---|---|
| Stage C (no scenario):<br>7. I worry about my privacy when I use audio technologies (audio assistance) but I feel that there is no point in trying to protect myself. Companies are more powerful. (Q143) | [Strongly disagree (5), Somewhat disagree (4), Neither agree nor disagree (3), Somewhat agree (2), Strongly agree (1)] |
| Scenario F1 – Internet:<br>8. Do you know what steps to take to protect your voice and personal data? (Q156) | [Definitely not (1), Probably not (2), Might or might not (3), Probably yes (4), Definitely yes (5)] |
| Scenario D3 – Home:<br>9. Do you know if your voice data is stored and analysed by any of the companies that provide voice-activated technology?(Q157) | [Yes (5), No (1)] |
| Scenario D3 – Home:<br>10. I know all about the risks that relate to the use of voice assistants. Companies' employees can hear my voice. (Q169) | [Not at all (1), Occasionally (2), Somewhat (3), Mostly (4), Definitely (5)] |
| Scenario D2 – Smart Environments:<br>11. If I knew my workplace was equipped with microphones, I would behave differently at work, including not talking with certain people or avoiding certain topics. (Q179) | [Strongly disagree (1), Somewhat disagree (2), Neither agree nor disagree (3), Somewhat agree (4), Strongly agree (5)] |
| Scenario E1 – TV/Film Industry Translation:<br>12. If I disagree with a translation, I know how to contact someone to express my complaints. (Q190) | [Strongly disagree (1), Somewhat disagree (2), Neither agree nor disagree (3), Somewhat agree (4), Strongly agree (5)] |
| Scenario F1 – Voice Conferencing:<br>13. I know what security protections are in place when using video conferencing tools for high-stakes business discussions. (Q194) | [Strongly disagree (1), Somewhat disagree (2), Neither agree nor disagree (3), Somewhat agree (4), Strongly agree (5)] |

**Table B.8**
Survey prompts for estimating perceived responsibility (PR).

| Survey Prompts | Responses |
|---|---|
| Stage C (no scenario):<br>1. Do you take precautions to protect other people's rights for example by getting permission when recording others and only using authorised material for work? (Q140) | [Never (1), Not often(2), Sometimes (3), Often(4), Always (5)] |
| Stage C (no scenario):<br>2. Do you take permission before you record someone's voice and conversations with others? (Q147) | [Yes (5), No (1)] |
| Stage C (no scenario):<br>3. Do you read the terms of use of the audio products you purchase to find out your rights and obligations (i.e. copy and share audio files)? (Q148) | [Yes (5), No (1)] |
| Stage C (no scenario):<br>4. Before you share copyright audio files or use them, do you search if it allowed by a licence? (Q149) | [Yes (5), No (1)] |

**Table B.9**
Survey Prompts for Estimating Exposure to Risks (EtR).

| Survey Prompts | Responses |
|---|---|
| Scenario C1 – Preserving Voice Memory of Loved Ones:<br>1. How concerned are you about the potential misuse of your loved one's preserved voice and personal stories? (Q118) | [Very concerned (5) , Somewhat concerned (4), Neutral (3), Not very concerned (2), Not at all concerned (1)] |
| Scenario C1 – Preserving Voice Memory of Loved Ones:<br>2. How comfortable are you with the idea of your loved one's voice and personal stories being used for research or commercial purposes? (Q122) | [Very comfortable (1), Somewhat comfortable (2), Neutral (3), Somewhat uncomfortable (4), Very uncomfortable (5)] |
| Scenario C1 – Preserving Voice Memory of Loved Ones:<br>3. How reluctant would you feel that your memories could be misinterpreted by future generations? (Q125) | [Very likely (5), Somewhat likely (4), Neutral (3), Somewhat unlikely (2), Very unlikely (1)] |

(*continued on next page*)

**Table B.9** (*continued*).

| Survey Prompts | Responses |
|---|---|
| Scenario E1 – Translation Services:<br>4. Automatically translated television and films may contain inappropriate content that is not suitable for children (even if the original language content was appropriate). (Q188) | [Strongly disagree (1), Somewhat disagree (2), Neither agree nor disagree (3), Somewhat agree (4), Strongly agree (5)] |
| Scenario E1 – Translation Services:<br>5. If I use a device that allows me to communicate in other languages with another person (e.g., a smartphone app), any sensitive things that I say may be stored in a database and someone can use that information without my consent. (Q189) | [Strongly disagree (1), Somewhat disagree (2), Neither agree nor disagree (3), Somewhat agree (4), Strongly agree (5)] |
| Scenario F1 – Work:<br>6. There are certain topics that I try to avoid when using video conferencing tools. (Q195) | [Strongly disagree (1), Somewhat disagree (2), Neither agree nor disagree (3), Somewhat agree (4), Strongly agree (5)] |
| Scenario F1 – Work:<br>7. My conversations are susceptible to eavesdropping. (Q196) | [Strongly disagree (1), Somewhat disagree (2), Neither agree nor disagree (3), Somewhat agree (4), Strongly agree (5)] |
| Scenario D3 – Home:<br>8. I would accept additional layers of security to protect my voice and audio data, even if they were inconvenient or annoying. (Q166) | [Strongly disagree (1), Somewhat disagree (2), Neither agree nor disagree (3), Somewhat agree (4), Strongly agree (5)] |
| Scenario D1 – Vehicle:<br>9. I believe I already have all of the privacy protections that I need to continue using audio technologies safely. (Q168) | [Not at all (5), Occasionally (4), Somewhat (3), Mostly (2), Definitely (1)] |
| Scenario G1 – AI Dolls:<br>10. How concerned are you about the privacy and security implications of an internet-connected doll that remembers your child's conversations and play history? (Q87) | [Very concerned(5), Somewhat concerned(4), Not concerned(1)] |
| Scenario G1 – AI Dolls:<br>11. Would you be okay with the internet-connected doll collecting data on your child's play patterns and sharing it with third-party companies for marketing or other purposes? (Q90) | [Yes(1), No(5), It depends on how the data is being used and who it is being shared with (3).] |
| Scenario G1 – AI Dolls:<br>12. How concerned are you about the potential for the internet-connected doll to be hacked or otherwise compromised, putting your child's privacy and security at risk? (Q92) | [Very concerned (5), Somewhat concerned(3), Not concerned (1)] |
| Scenario D2 – Work:<br>13. The microphones that my employer would install in my workplace are meant to spy on me and ultimately can be used to get me fired. (Q178) | [Strongly disagree (1), somewhat disagree (2), neither agree nor disagree (3), somewhat agree (4)] |

**Table B.10**
Survey prompts for estimating perceived trust (PT).

| Survey Prompts | Responses |
|---|---|
| Scenario A1 – Smart Hearing Aid:<br>1. How do you think the use of such a hearing aid affects the dynamics of interpersonal communication and trust? (Q235) | [Positively(5), Negatively(1), Not sure(3)] |
| Scenario A2 – AI-Assistive Device for the Blind:<br>2. How much do you trust AI-assistive devices to accurately describe surroundings? (Q242) | [Completely trust (5), Somewhat trust (3), Do not trust at all(1)] |
| Scenario H2 – Social Interactive Robot for Nursing:<br>3. Would you trust a social-interactive robot to provide essential medical reminders to your elderly relative? (Q222) | [Yes (5), No (1), Neutral (3)] |
| Scenario H2 – Social Interactive Robot for Nursing:<br>4. Would you trust a social-interactive robot to provide comforting support for loneliness or hold a casual conversation with your elderly relative? (Q223) | [Yes (5), No (1), Neutral (3)] |
| Scenario H1 – Health Monitoring App:<br>5. Do you trust the accuracy and reliability of an AI-based app for monitoring your breathing disorder? (Q212) | [Yes (5), No (1), Not sure (3)] |

(*continued on next page*)

**Table B.10** (*continued*).

| Survey Prompts | Responses |
|---|---|
| Scenario I1 – Live Singing:<br>6. Do you trust the technology used to capture and reproduce your voice for this performance? (Q127) | [Yes(5), No(1), Unsure(3)] |
| Scenario I1 – Live Singing:<br>7. Would you trust this system more if the content was deleted after the performance. (Q128) | [Yes(5), No(1), Unsure(3)] |
| Scenario I2 – Pets:<br>8. Do you trust that a system like this would accurately convey your intended message to your pet? (Q274) | [Yes, I trust that the system would accurately convey my intended message. (5), No, I don't trust that the system would accurately convey my intended message.(1), Unsure.(3)] |
| Scenario B1 – Creating Audio Books:<br>9. Do you trust these technologies? Believe they function legally, ethically and responsibly? (Q304) | [Yes (5), No (1), I don't know(3)] |
| Scenario D1 – Vehicle:<br>10. I don't trust a voice assistant in my car with purchasing decisions made on my behalf, even if the suggested products are things that I need. (Q172) | Never (5) Not often (4) Sometimes (3) Often or always (1) |
| Scenario D3 – Home:<br>11. I trust one particular voice assistant brand over another because of my concerns about technology that "always listens". (Q164) | [Yes (5),No (1)] |
| Scenario G1 – AI Dolls:<br>12. Do you trust the company that makes the internet-connected doll to handle your child's data and information responsibly? (Q88) | [Yes, I trust them (5), No, I do not trust them(1), I'm not sure(3)] |

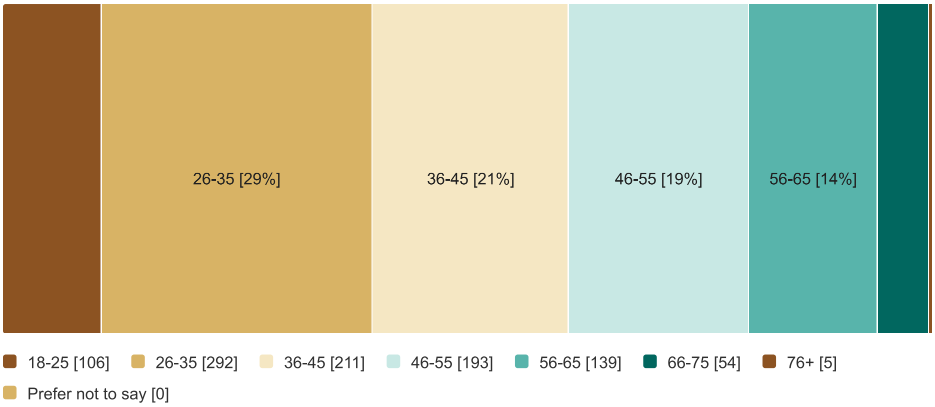
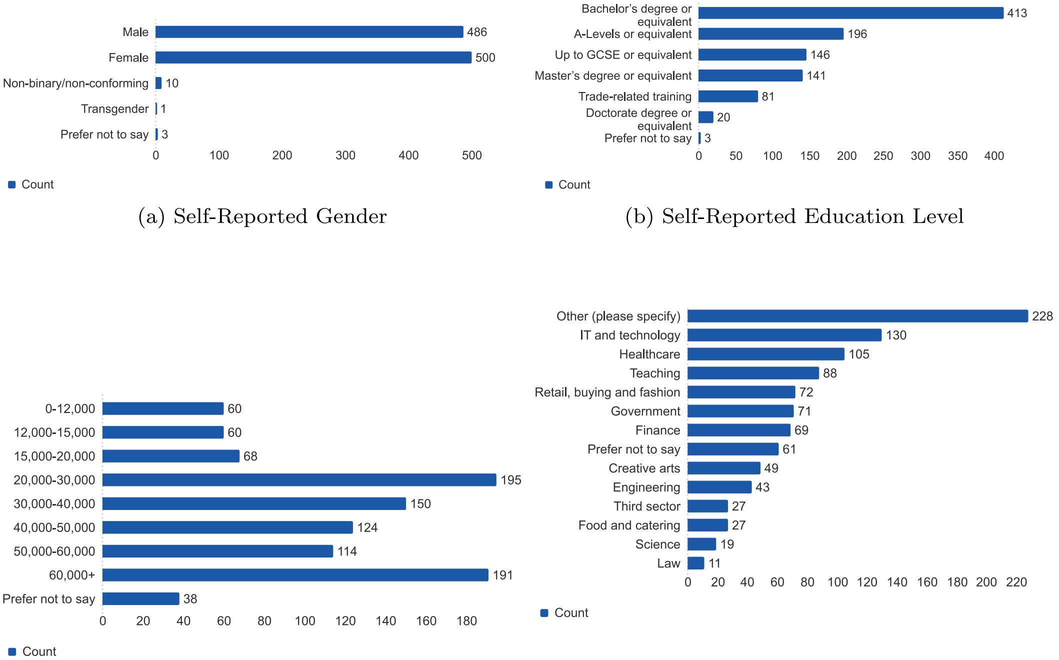


**Fig. C.3.** Participant age category.



(a) Self-Reported Gender

(b) Self-Reported Education Level

(a) Self-Reported Income (UK GBP)

(b) Self-Reported Occupation

**Fig. C.4.** Self-reported demographics for gender, education level, income, and occupation.

**Table D.11**
One way ANOVA effect of gender on perceived awareness of rights (PAoR)

(a)One-way ANOVA for Gender and PAoR ($p = 0.001$, two-tailed).

| | **Sum of Squares** | **df** | **Mean Square** | **F** | **Sigma** |
|---|---|---|---|---|---|
| **Between Groups** | 0.97 | 4 | 0.24 | 8.09 | $< 0.001$ |
| **Within Groups** | 29.88 | 995 | 0.03 | | 0.15 |
| **Total** | 30.85 | 999 | | | |

(b) One-way ANOVA Effect Size of Gender on Perceived Awareness of Rights (PAoR). Eta square ($\eta^2$) is .032 (95% CI [0.011, 0.052]), indicating a statistically significant but small effect size (F= 8.09, p = 0.001) of gender on perceived awareness of rights.

| | | | **95% Confidence Interval** | |
|---|---|---|---|---|
| | | **Point Estimate** | **Lower** | **Upper** |
| **PAoR** | Eta-Squared | 0.032 | 0.011 | 0.052 |
| | Epsilon Squared | 0.028 | 0.007 | 0.049 |
| | Omega-Squared Fixed Effect | 0.028 | 0.007 | 0.049 |
| | Omega-Squared Random Effect | 0.007 | 0.002 | 0.013 |

**Table D.12**
One Way ANOVA Effect of Gender on Perceived Trust (PT).

(a)One-way ANOVA for Gender and PT ($p = 0.001$, two-tailed).

| | Sum of Squares | df | Mean Square | F | Sigma |
|---|---|---|---|---|---|
| **Between Groups** | 1.03 | 4 | 0.257 | 8.317 | < 0.001 |
| **Within Groups** | 30.801 | 995 | 0.031 | | |
| **Total** | 31.831 | 999 | | | |

(b) One-way ANOVA Effect Size of Gender on Perceived Trust (PT). Eta square ($\eta^2$) is .032 (95% CI [0.012, 0.053]), indicating a statistically significant but small effect size (F= 8.31, p = 0.001) of gender on perceived trust.

| | | | 95% Confidence Interval | |
|---|---|---|---|---|
| | | **Point Estimate** | **Lower** | **Upper** |
| **PT** | Eta-Squared | 0.032 | 0.012 | 0.053 |
| | Epsilon Squared | 0.028 | 0.008 | 0.050 |
| | Omega-Squared Fixed Effect | 0.028 | 0.008 | 0.050 |
| | Omega-Squared Random Effect | 0.007 | 0.002 | 0.013 |

**Table D.13**
Statistical summary of Pearson correlation, sum of squares and covariance for PAoR and PT. **The correlation is significant (p = 0.01, two-tailed.)

| | Correlations | | |
|---|---|---|---|
| | | **Perceived Trust (PT)** | **Perceived Awareness of Rights (PAoR)** |
| **Perceived Trust (PT)** | Pearson Correlation | 1 | 0.13** |
| | Sig(2-tailed) | | < 0.001 |
| | Sum of Squares and Cross Products | 548.02 | 55.61 |
| | Covariance | 0.54 | 0.05 |
| | N | 1000 | 1000 |
| **Perceived Awareness of Rights (PAoR)** | Pearson Correlation | 0.13** | 1 |
| | Sig(2-tailed) | < 0.001 | |
| | Sum of Squares and Cross Products | 55.61 | 316.69 |
| | Covariance | 0.05 | 0.31 |
| | N | 1000 | 1000 |

**Table D.14**
Statistical summary of Pearson correlation, sum of squares and covariance for PR and EtR. **The correlation is significant (p = 0.01, two-tailed.)

| | Correlations | | |
|---|---|---|---|
| | | **Perceived Responsibility (PR)** | **Exposure to Risks (EtR)** |
| **Perceived Responsibility (PR)** | Pearson Correlation | 1 | 0.14** |
| | Sig(2-tailed) | | < 0.001 |
| | Sum of Squares and Cross Products | 1438.09 | 87.34 |
| | Covariance | 1.44 | 0.08 |
| | N | 1000 | 1000 |
| **Exposure to Risks (EtR)** | Pearson Correlation | 0.14** | 1 |
| | Sig(2-tailed) | < 0.001 | |
| | Sum of Squares and Cross Products | 87.34 | 271.88 |
| | Covariance | 0.08 | 0.27 |
| | N | 1000 | 1000 |

**Table D.15**
Statistical summary of Pearson correlation, sum of squares and covariance for PAoR and PR. **The correlation is significant (p = 0.01, two-tailed.)

| | Correlations | | |
|---|---|---|---|
| | | **Perceived Awareness of Rights (PAoR)** | **Perceived Responsibility (PR)** |
| **Perceived Awareness of Rights (PAoR)** | Pearson Correlation | −− | |
| | Sum of Squares and Cross Products | 316.691 | |
| | Covariance | 0.31 | |
| | N | 1000 | |
| **Perceived Responsibility (PR)** | Pearson Correlation | 0.46** | −− |
| | Sig(2-tailed) | < 0.001 | |
| | Sum of Squares and Cross Products | 311.45 | 1438.09 |
| | Covariance | 0.31 | 1.440 |
| | N | 1000 | 1000 |

**Table D.16**
Statistical summary of Pearson correlation, sum of squares and covariance for PT and EtR. **The correlation is significant (p = 0.01, two-tailed.)

| | Correlations | | |
|---|---|---|---|
| | | **Perceived Trust (PT)** | **Exposure to Risks (EtR)** |
| **Perceived Trust (PT)** | Pearson Correlation | −− | |
| | Sum of Squares and Cross Products | 548.024 | |
| | Covariance | 0.549 | |
| | N | 1000 | |
| **Exposure to Risks (EtR)** | Pearson Correlation | −0.491** | 1 |
| | Sig(2-tailed) | < 0.001 | |
| | Sum of Squares and Cross Products | -189.45 | 271.88 |
| | Covariance | -0.19 | -0.27 |
| | N | 1000 | 1000 |

**Table D.17**
One Way ANOVA Effect of Education on All Measured Survey Variables.

(a) One-way ANOVA of Education and Survey Variables (p = 0.001, two-tailed).

| | | Sum of Squares | df | Mean Square | F | Sigma |
|---|---|---|---|---|---|---|
| PR | **Between Groups** | 14.849 | 5 | 2.970 | 2.070 | 0.067 |
| | **Within Groups** | 1422.064 | 991 | 1.435 | | |
| | **Total** | 1436.912 | 996 | | | |
| PAoR | **Between Groups** | 3.449 | 5 | 0.690 | 2.183 | 0.054 |
| | **Within Groups** | 313.094 | 991 | 0.316 | | |
| | **Total** | 316.543 | 996 | | | |
| PT | **Between Groups** | 1.867 | 5 | 0.373 | 0.677 | 0.641 |
| | **Within Groups** | 546.150 | 991 | | | |
| | **Total** | 548.017 | 996 | | | |
| EtR | **Between Groups** | 1.150 | 5 | 0.230 | 0.843 | 0.519 |
| | **Within Groups** | 270.425 | 991 | | | |
| | **Total** | 271.575 | 996 | | | |

(b) One-way ANOVA Effect Size of Education on all the Survey Variables. The demographic variable Education is not found to be statistically significant to any of the four Survey Variables (PR, PAoR, PT, and EtR).

| | | | 95% Confidence Interval | |
|---|---|---|---|---|
| | | Point Estimate | Lower | Upper |
| **PR** | Eta-Squared | 0.010 | 0.000 | 0.021 |
| | Epsilon Squared | 0.005 | -0.005 | 0.016 |
| | Omega-Squared Fixed Effect | 0.005 | -0.005 | 0.016 |
| | Omega-Squared Random Effect | 0.001 | -0.001 | 0.003 |
| **PAoR** | Eta-Squared | 0.011 | 0.000 | 0.022 |
| | Epsilon Squared | 0.006 | -0.005 | 0.017 |
| | Omega-Squared Fixed Effect | 0.006 | -0.005 | 0.017 |
| | Omega-Squared Random Effect | 0.001 | -0.001 | 0.004 |
| **PT** | Eta-Squared | 0.003 | 0.000 | 0.008 |
| | Epsilon Squared | -0.002 | -0.005 | 0.003 |
| | Omega-Squared Fixed Effect | -0.002 | -0.005 | 0.003 |
| | Omega-Squared Random Effect | 0.000 | -0.001 | 0.001 |
| **EtR** | Eta-Squared | 0.004 | 0.000 | 0.010 |
| | Epsilon Squared | -0.001 | -0.005 | 0.005 |
| | Omega-Squared Fixed Effect | -0.001 | -0.005 | 0.005 |
| | Omega-Squared Random Effect | 0.000 | -0.001 | 0.001 |

**Table D.18**

One Way ANOVA Effect of Age on All Measured Survey Variables.

(a) One-way ANOVA for Age with each Survey Variable (p = 0.001, two-tailed).

| | | Sum of Squares | df | Mean Square | F | Sigma |
|---|---|---|---|---|---|---|
| PR | **Between Groups** | 2.379 | 6 | 0.397 | 0.274 | 0.949 |
| | **Within Groups** | 1435.718 | 993 | 1.446 | | |
| | **Total** | 1438.097 | 999 | | | |
| PAoR | **Between Groups** | 1.899 | 6 | 0.316 | 0.998 | 0.425 |
| | **Within Groups** | 314.792 | 993 | 0.317 | | |
| | **Total** | 316.691 | 999 | | | |
| PT | **Between Groups** | 3.543 | 6 | 0.591 | 1.077 | 0.374 |
| | **Within Groups** | 544.481 | 993 | 0.548 | | |
| | **Total** | 548.024 | 999 | | | |
| EtR | **Between Groups** | 2.544 | 6 | 0.424 | 1.563 | 0.155 |
| | **Within Groups** | 269.340 | 993 | 0.271 | | |
| | **Total** | 271.884 | 999 | | | |

(b)One-way ANOVA Effect Size of Age on all the Survey Variables. The demographic variable Age is not found to be statistically significant to any of the four Survey Variables (PR, PAoR, PT, and EtR).

| | | | 95% Confidence Interval | |
|---|---|---|---|---|
| | | Point Estimate | Lower | Upper |
| **PR** | Eta-Squared | 0.002 | 0.000 | 0.002 |
| | Epsilon Squared | -0.004 | -0.006 | -0.004 |
| | Omega-Squared Fixed Effect | -0.004 | -0.006 | -0.004 |
| | Omega-Squared Random Effect | -0.001 | -0.001 | -0.001 |
| **PAoR** | Eta-Squared | 0.006 | 0.000 | 0.012 |
| | Epsilon Squared | 0.000 | -0.006 | 0.006 |
| | Omega-Squared Fixed Effect | 0.000 | -0.006 | 0.006 |
| | Omega-Squared Random Effect | 0.000 | -0.001 | 0.001 |
| **PT** | Eta-Squared | 0.006 | 0.000 | 0.013 |
| | Epsilon Squared | 0.000 | -0.006 | 0.007 |
| | Omega-Squared Fixed Effect | 0.000 | -0.006 | 0.007 |
| | Omega-Squared Random Effect | 0.000 | -0.001 | 0.001 |
| **EtR** | Eta-Squared | 0.009 | 0.000 | 0.018 |
| | Epsilon Squared | 0.003 | -0.006 | 0.013 |
| | Omega-Squared Fixed Effect | 0.003 | -0.006 | 0.013 |
| | Omega-Squared Random Effect | 0.001 | -0.001 | 0.002 |

## Data availability

We include a “Data Availability Statement” in the paper itself, with a persistent link to the dataset with DOI hosted by Zenodo.

## References

Aithal, Architha, Aithal, P.S., 2020. Development and validation of survey questionnaire & experimental data–a systematical review-based statistical approach. Int. J. Manag. Technol. Soc. Sciences (IJMTS) 5 (2), 233–251.

Alharbi, Sadeen, Alrazgan, Muna, Alrashed, Alanoud, Alnomasi, Turkiayh, Almojel, Raghad, Alharbi, Rimah, Alharbi, Saja, Alturki, Sahar, Alshehri, Fatimah, Almojil, Maha, 2021. Automatic speech recognition: Systematic literature review. IEEE Access 9, 131858–131876. http://dx.doi.org/10.1109/ACCESS.2021.3112535.

Ameen, Nisreen, Sharma, Gagan Deep, Tarba, Shlomo, Rao, Amar, Chopra, Ritika, 2022. Toward advancing theory on creativity in marketing and artificial intelligence. Psychol. Mark. 39 (9), 1802–1825.

An, KwangHoon, Kim, Myung Jong, Teplansky, Kristin, Green, Jordan R, Campbell, Thomas F, Yunusova, Yana, Heitzman, Daragh, Wang, Jun, 2018. Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks. In: Proceedings of Interspeech. pp. 1913–1917.

Austin, Michael L., 2019. Is siri a little bit racist? Recognizing and confronting algorithmic bias in emerging media. In: Race/Gender/Class/Media. Routledge, pp. 246–250.

Babaev, Nicholas, Tamogashev, Kirill, Saginbaev, Azat, Shchekotov, Ivan, Bae, Hanbin, Sung, Hosang, Lee, WonJun, Cho, Hoon-Young, Andreev, Pavel, 2024. FINALLY: fast and universal speech enhancement with studio-like quality. Adv. Neural Inf. Process. Syst. 37, 934–965.

Baevski, Alexei, Zhou, Henry, Mohamed, Abdelrahman, Auli, Michael, 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. 33, 12449–12460.

Bajorek, Joan Palmiter, 2019. Voice recognition still has significant race and gender biases. Harv. Bus. Rev. 10, 1–4.

Bengio, Yoshua, Mindermann, Sören, Privitera, Daniel, Besiroglu, Tamay, Bommasani, Rishi, Casper, Stephen, Choi, Yejin, Fox, Philip, Garfinkel, Ben, Goldfarb, Danielle, et al., 2025. International AI safety report. arXiv preprint arXiv:2501.17805.

Bommasani, Rishi, Hudson, Drew A, Adeli, Ehsan, Altman, Russ, Arora, Simran, von Arx, Sydney, Bernstein, Michael S, Bohg, Jeannette, Bosselut, Antoine, Brunskill, Emma, et al., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Boone Jr., Harry N., Boone, Deborah A., 2012. Analyzing likert data. J. Ext. 50 (2), 48.

Campbell, Joseph P., 1997. Speaker recognition: A tutorial. Proc. the IEEE 85 (9), 1437–1462.

Capraro, Valerio, Lentsch, Austin, Acemoglu, Daron, Akgun, Selin, Akhmedova, Aisel, Bilancini, Ennio, Bonnefon, Jean-François, Brañas-Garza, Pablo, Butera, Luigi, Douglas, Karen M, et al., 2024. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS Nexus 3 (6), pgae191.

Chakrabartty, Satyendra Nath, 2020. Combining likert items with different number of response categories. Proc. Eng. Sci. 2 (3), 311–322.

Chan, William, Jaitly, Navdeep, Le, Quoc, Vinyals, Oriol, 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4960–4964.

Chen, Zehua, Tan, Xu, Wang, Ke, Pan, Shifeng, Mandic, Danilo, He, Lei, Zhao, Sheng, 2022. Infergrad: Improving diffusion models for vocoder by considering inference in training. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 8432–8436.

Cheng, Ho Kei, Oh, Seoung Wug, Price, Brian, Schwing, Alexander, Lee, Joon-Young, 2023. Tracking anything with decoupled video segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1316–1326.

Converse, Jean M., Presser, Stanley, 1986. Survey Questions: Handcrafting the standardized questionnaire. vol. 63, Sage.

Dang, Trung, Tran, Dung, Chin, Peter, Koishida, Kazuhito, 2022. Training robust zero-shot voice conversion models with self-supervised features. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6557–6561.

Datta, Prerit, Namin, Akbar Siami, Chatterjee, Moitrayee, 2018. A survey of privacy concerns in wearable devices. In: IEEE International Conference on Big Data (Big Data). pp. 4549–4553.

DCASE, 2025. Detection and Classification of Acoustic Scenes and Events. https://dcase.community/community_info. (Accessed: 2025-03-31).

Delmonte, Rodolfo, Mian, G., Tisato, Graziano, 1986. A grammatical component for a text-to-speech system. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. 11, pp. 2407–2410.

Dhiya'Mardhiyyah, Alya, Latif, Jazlyn Jan Keyla, Tho, Cuk, 2023. Privacy and security in the use of voice assistant: An evaluation of user awareness and preferences. In: IEEE International Conference on Information Management and Technology (ICIMTech). pp. 481–486.

Fan, Yuchen, Qian, Yao, Soong, Frank K., He, Lei, 2015. Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4475–4479.

Förster, Frank, Romeo, Marta, Holthaus, Patrick, Wood, Luke J, Dondrup, Christian, Fischer, Joel E, Liza, Farhana Ferdousi, Kaszuba, Sara, Hough, Julian, Nesset, Birthe, et al., 2023. Working with roubles and failures in conversation between humans and robots: Workshop report. Front. Robot. AI 10, 1202306.

Ghasemi, Asghar, Zahediasl, Saleh, 2012. Normality tests for statistical analysis: A guide for non-statisticians. Int. JOurnal ENdocrinology MEtabolism 10 (2), 486.

Glass, Gene V., Peckham, Percy D., Sanders, James R., 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev. Educ. Res. 42 (3), 237–288.

Google, 2025. Project Astra. https://deepmind.google/models/project-astra. (Accessed: 2025-03-31).

Graves, Alex, Fernández, Santiago, Gomez, Faustino, Schmidhuber, Jürgen, 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: International Conference on Machine Learning. ICML '06, pp. 369–376. http://dx.doi.org/10.1145/1143844.1143891, URL http://dx.doi.org/10.1145/1143844.1143891.

Graves, Alex, Mohamed, Abdel-rahman, Hinton, Geoffrey, 2013. Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6645–6649.

Gulati, Anmol, Qin, James, Chiu, Chung-Cheng, Parmar, Niki, Zhang, Yu, Yu, Jiahui, Han, Wei, Wang, Shibo, Zhang, Zhengdong, Wu, Yonghui, Pang, Ruoming, 2020. Conformer: Convolution-augmented transformer for speech recognition. In: Proceedings of Interspeech. pp. 5036–5040. http://dx.doi.org/10.21437/Interspeech.2020-3015.

Gupta, Deepak, Attal, Kush, Demner-Fushman, Dina, 2023. A dataset for medical instructional video classification and question answering. Sci. Data 10 (1), 158.

Gyevnár, Bálint, Kasirzadeh, Atoosa, 2025. AI safety for everyone. Nat. Mach. Intell. 1–12.

Hair Jr., J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M., 2013. A primer on partial least squares structural equation modeling (PLS-SEM). Eur. J. Tour. Res. 6 (2), 211–213.

Harel, Brian T, Cannizzaro, Michael S, Cohen, Henri, Reilly, Nicole, Snyder, Peter J, 2004. Acoustic characteristics of Parkinsonian speech: A potential biomarker of early disease progression and treatment. J. Neurolinguistics 17 (6), 439–453.

Harwell, Michael R., 1992. Summarizing Monte Carlo results in methodological research. J. Educ. Stat. 17 (4), 297–313.

Hawley, Janet L., Hancock, Adrienne B., 2024. Incorporating mobile app technology in voice modification protocol for transgender women. J. Voice 38 (2), 337–345.

Hirschberg, Julia, 2006. Speech synthesis: Prosody. Encycl. Lang. Linguist. 7, 49–55.

Hunt, Andrew J., Black, Alan W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. 1, pp. 373–376.

Hutiri, Wiebke, Papakyriakopoulos, Orestis, Xiang, Alice, 2024. Not my voice! a taxonomy of ethical and safety harms of speech generators. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 359–376.

IBM Corp, 2022. IBM SPSS Statistics for Windows. Armonk, NY: IBM Corp, version 29.0, https://hadoop.apache.org.

Kalchbrenner, Nal, Elsen, Erich, Simonyan, Karen, Noury, Seb, Casagrande, Norman, Lockhart, Edward, Stimberg, Florian, Oord, Aaron, Dieleman, Sander, Kavukcuoglu, Koray, 2018. Efficient neural audio synthesis. In: International Conference on Machine Learning. PMLR, pp. 2410–2419.

Khandelwal, Tanmay, Das, Rohan Kumar, Chng, Eng Siong, et al., 2024. Sound event detection: A journey through DCASE challenge series. APSIPA Trans. Signal Inf. Process. 13 (1).

Kim, Seoyoung, Park, Yeon Su, Ahn, Dakyeom, Kwak, Jin Myung, Kim, Juho, 2024. Is the same performance really the same?: Understanding how listeners perceive ASR results differently according to the speaker's accent. Proc. the ACM Human-Computer Interact. 8 (CSCW1), 1–22.

Kinnunen, Tomi, Sahidullah, Md., Delgado, Héctor, Todisco, Massimiliano, Evans, Nicholas, Yamagishi, Junichi, Lee, Kong Aik, 2017. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In: Proceedings of Interspeech. pp. 2–6.

Koelle, Marion, Ananthanarayan, Swamy, Czupalla, Simon, Heuten, Wilko, Boll, Susanne, 2018. Your smart glasses' camera bothers me! exploring opt-in and opt-out gestures for privacy mediation. In: Proceedings of the 10th Nordic Conference on Human-Computer Interaction. pp. 473–481.

Koenecke, Allison, Nam, Andrew, Lake, Emily, Nudell, Joe, Quartey, Minnie, Mengesha, Zion, Toups, Connor, Rickford, John R., Jurafsky, Dan, Goel, Sharad, 2020. Racial disparities in automated speech recognition. Proc. Natl. Acad. Sci. 117 (14), 7684–7689. http://dx.doi.org/10.1073/pnas.1915768117, URL https://www.pnas.org/doi/10.1073/pnas.1915768117.

Kong, Jungil, Kim, Jaehyeon, Bae, Jaekyoung, 2020. Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. Adv. Neural Inf. Process. Syst. 33, 17022–17033.

Kowalczuk, Pascal, 2018. Consumer acceptance of smart speakers: a mixed methods approach. J. Res. Interact. Mark. 12 (4), 418–431. http://dx.doi.org/10.1108/JRIM-01-2018-0022, URL http://dx.doi.org/10.1108/JRIM-01-2018-0022.

Kulkarni, Pranav, Duffy, Orla, Synnott, Jonathan, Kernohan, W. George, McNaney, Roisin, 2022. Speech and language practitioners' experiences of commercially available voice-assisted technology: Web-based survey study. JMIR Rehabil. Assist. Technol. 9 (1), e29249. http://dx.doi.org/10.2196/29249, URL https://rehab.jmir.org/2022/1/e29249.

Lau, Josephine, Zimmerman, Benjamin, Schaub, Florian, 2018. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. Proc. ACM Human-Computer Interact. 2 (CSCW), 102:1–102:31. http://dx.doi.org/10.1145/3274371, URL https://dl.acm.org/doi/10.1145/3274371.

Laux, Johann, Wachter, Sandra, Mittelstadt, Brent, 2024. Trustworthy artificial intelligence and the European union AI act: On the conflation of trustworthiness and acceptability of risk. Regul. Gov. 18 (1), 3–32.

Leschanowsky, Anna, Rech, Silas, Popp, Birgit, Bäckström, Tom, 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, URL https://www.sciencedirect.com/science/article/pii/S0747563224002127.

Lian, Jiachen, Zhang, Chunlei, Yu, Dong, 2022. Robust disentangled variational speech representation learning for zero-shot voice conversion. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6572–6576.

Likert, Rensis, 1932. A technique for the measurement of attitudes. Arch. Psychol..

Lin, Ji, Gan, Chuang, Han, Song, 2019. TSM: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7083–7093.

Liu, Min, Li, Rita Yi Man, Deeprasert, Jirawan, 2024. Factors that affect individuals in using digital currency electronic payment in China: SEM and fsQCA approaches. Int. Rev. Econ. Financ. 95, 103418.

Lix, Lisa M., Keselman, Joanne C., Keselman, Harvey J., 1996. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Rev. Educ. Res. 66 (4), 579–619.

Lüdemann, René, Schulz, Alexander, Kuhl, Ulrike, 2024. Generation gap or diffusion trap? How age affects the detection of personalized AI-generated images. In: International Conference on Computer-Human Interaction Research and Applications. Springer, pp. 359–381.

MIRO, 2025. Miro online whiteboard (no version provided). https://miro.com. (Accessed: 2025-03-31).

Mitra, Vikramjit, Shriberg, Elizabeth, 2015. Effects of feature type, learning algorithm and speaking style for depression detection from speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4774–4778.

Mohammad, Ahmad Saeed, Al-Ani, Jabir Alshehabi, 2017. Towards ethnicity detection using learning based classifiers. In: IEEE Proceedings of Computer Science and Electronic Engineering. pp. 219–224.

Müller, Nicolas M., Kawa, Piotr, Hu, Shen, Neu, Matthias, Williams, Jennifer, Sperl, Philip, Böttinger, Konstantin, 2024. A new approach to voice authenticity. In: Interspeech 2024. pp. 2245–2249. http://dx.doi.org/10.21437/Interspeech.2024-31.

Nacimiento-García, Eduardo, Díaz-Kaas-Nielsen, Holi Sunya, González-González, Carina S, 2024. Gender and accent biases in AI-based tools for spanish: A comparative study between alexa and whisper. Appl. Sci. 14 (11), 4734.

Nandhini, T.J., Thinakaran, K., 2023. SIFT algorithm-based object detection and tracking in the video image. In: Fifth International Conference on Electrical, Computer and Communication Technologies. pp. 1–4.

Nautsch, Andreas, Jasserand, Catherine, Kindt, Els, Todisco, Massimiliano, Trancoso, Isabel, Evans, Nicholas, 2019. The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding. In: Proceedings of Interspeech. pp. 3695–3699. http://dx.doi.org/10.21437/Interspeech.2019-2647.

Newman, Michael, Wu, Angela, 2011. "Do you sound Asian when you speak english?" racial identification and voice in Chinese and Korean Americans' english. J. Am. Speech 86 (2), 152–178.

Ntalampiras, Stavros, Potamitis, Ilyas, Fakotakis, Nikos, 2012. Acoustic detection of human activities in natural environments. J. the Audio Eng. Soc. 60 (9), 686–695.

Nuthakki, Ramesh, Masanta, Payel, Yukta, T., 2022. A Literature Survey on Speech Enhancement Based on Deep Neural Network Technique. Springer, pp. 7–16. http://dx.doi.org/10.1007/978-981-16-7985-8_2.

OpenAI, 2025. Open AI Chat GPT. https://openai.com/index/chatgpt. (Accessed: 2025-03-31).

Owen, Richard, Stilgoe, Jack, Macnaghten, Phil, Gorman, Mike, Fisher, Erik, Guston, Dave, 2013. A framework for responsible innovation. Responsible Innov.: Manag. Responsible Émerg. of Sci. Innov. Soc. 27–50.

Pal, Debajyoti, Arpnikanondt, Chonlameth, Funilkul, Suree, Varadarajan, Vijayakumar, 2019. User experience with smart voice assistants: The accent perspective. In: IEEE International Conference on Computing, Communication and Networking Technologies. pp. 1–6.

Picard, Christopher, Smith, Katherine Elizabeth, Picard, Kelly, Douma, Matthew John, 2020. Can alexa, cortana, google assistant and siri save your life? A mixed-methods analysis of virtual digital assistants and their responses to first aid and basic life support queries. Br. Med. J. Innov. 6 (1).

Portillo, Virginia, Greenhalgh, Chris, Craigon, Peter J, Ten Holter, Carolyn, 2023. Responsible research and innovation (RRI) prompts and practice cards: A tool to support responsible practice. In: Proceedings of the First International Symposium on Trustworthy Autonomous Systems. pp. 1–4.

Pradhan, Alisha, Mehta, Kanika, Findlater, Leah, 2018. "Accessibility came by accident": Use of voice-controlled intelligent personal assistants by people with disabilities. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, New York, NY, USA, pp. 1–13. http://dx.doi.org/10.1145/3173574.3174033, URL https://dl.acm.org/doi/10.1145/3173574.3174033.

Radford, Alec, Kim, Jong Wook, Xu, Tao, Brockman, Greg, McLeavey, Christine, Sutskever, Ilya, 2023. Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. PMLR, pp. 28492–28518.

Ren, Yi, Liu, Jinglin, Tan, Xu, Zhang, Chen, Qin, Tao, Zhao, Zhou, Liu, Tie-Yan, 2020. SimulSpeech: End-to-end simultaneous speech to text translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3787–3796.

Roekhaut, Sophie, Goldman, Jean-Philippe, Simon, Anne Catherine, et al., 2010. A model for varying speaking style in TTS systems. In: Proceedings of International Conference on Speech Prosody.

Romero, José Luis, Speckbacher, Michael, 2024. Estimation of binary time-frequency masks from ambient noise. SIAM J. Math. Anal. 56 (3), 3559–3587.

Saini, Parul, Kumar, Krishan, Kashid, Shamal, Saini, Ashray, Negi, Alok, 2023. Video summarization using deep learning techniques: A detailed analysis and investigation. Artif. Intell. Rev. 56 (11), 12347–12385.

Schröter, Hendrik, Rosenkranz, Tobias, Escalante-B, Alberto-N, Maier, Andreas, 2022. Low latency speech enhancement for hearing aids using deep filtering. IEEE/ACM Trans. Audio, Speech, Lang. Process. 30, 2716–2728.

Shan, Yue, Li, Rita Yi Man, 2025. The impact of AI recommendation's price and product accuracy on customer satisfaction: SEM-SOR theoretical approach. Curr. Psychol. 44 (8), 6978–6988.

Stensen, Kenneth, Lydersen, Stian, 2022. Internal consistency: from alpha to omega. Tidsskr. Den Nor. Laegeforening: Tidsskr. Praktisk Med. Ny Raekke 142 (12).

Stokes, Patrick, 2015. Deletion as second death: The moral status of digital remains. Ethics Inf. Technol. 17, 237–248.

Stypinska, Justyna, 2023. AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies. AI Soc. 38 (2), 665–677.

Su, Jiaqi, Jin, Zeyu, Finkelstein, Adam, 2021. HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features. In: 2021 IEEE Workshop on Applications of Signal Processing To Audio and Acoustics. WASPAA, IEEE, pp. 166–170.

Sweeney, Miriam, Davis, Emma, 2020. Alexa, are you listening?an exploration of smart voice assistant use and privacy in libraries. Inf. Technol. Libr. 39 (4).

Tatman, Rachael, 2017. Gender and dialect bias in YouTube's automatic captions. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Association for Computational Linguistics, Valencia, Spain, pp. 53–59. http://dx.doi.org/10.18653/v1/W17-1606, URL http://aclweb.org/anthology/W17-1606.

Todisco, Massimiliano, Wang, Xin, Vestman, Ville, Sahidullah, Md., Delgado, Héctor, Nautsch, Andreas, Yamagishi, Junichi, Evans, Nicholas, Kinnunen, Tomi H., Lee, Kong Aik, 2019. ASVspoof 2019: Future Horizons in spoofed and fake audio detection. In: Proceedings of Interspeech. pp. 1008–1012.

Tomashenko, Natalia, Srivastava, Brij Mohan Lal, Wang, Xin, Vincent, Emmanuel, Nautsch, Andreas, Yamagishi, Junichi, Evans, Nicholas, Patino, Jose, Bonastre, Jean-François, Noé, Paul-Gauthier, Todisco, M., 2020. Introducing the VoicePrivacy initiative. In: Proceedings of Interspeech. pp. 1693–1697.

UK Governmenet, 2025a. AI Safety Summit. https://www.aisafetysummit.gov.uk. (Accessed: 2025-03-31).

UK Government, 2023a. Ofcom, a deep dive into deepfakes that demean, defraud and disinform (2024). Ofcom Discuss. Pap. URL https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/deepfakes-demean-defraud-disinform.

UK Government, 2023b. A pro-innovation approach to AI regulation. Present. Parliam. By Secr. State Sci., Innov. Technol. By Command. His Majesty 29 March 2023 URL https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

UK Government, 2024. Implementing the UK's AI regulatory principles: Initial guidance for regulators. Dep. Sci. Innov. Technol. Foreign, Commonw. Dev. Off. Prime Minister's Off. 10 Downing Str. URL https://assets.publishing.service.gov.uk/media/65c0b6bd63a23d0013c821a0/implementing_the_uk_ai_regulatory_principles_guidance_for_regulators.pdf.

UK Government, 2025b. AI Seoul Summit 2024. https://www.gov.uk/government/topical-events/ai-seoul-summit-2024. (Accessed: 2025-03-31).

Urquhart, Lachlan, Miranda, Diana, Podoletz, Lena, 2022. Policing the smart home: The internet of things as 'invisible witnesses'. Inf. Polity 27 (2), 233–246.

van den Oord, Aäron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, Kavukcuoglu, Koray, 2016. WaveNet: A generative model for raw audio. In: Proceedings of 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9). p. 125.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, Polosukhin, Illia, 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Veale, Michael, Binns, Reuben, Ausloos, Jef, 2018. When data protection by design and data subject rights clash. Int. Data Priv. Law 8 (2), 105–123.

Wang, Xin, Delgado, Héctor, Tak, Hemlata, weon Jung, Jee, jin Shim, Hye, Todisco, Massimiliano, Kukanov, Ivan, Liu, Xuechen, Sahidullah, Md, Kinnunen, Tomi H., Evans, Nicholas, Lee, Kong Aik, Yamagishi, Junichi, 2024. ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In: The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024). pp. 1–8. http://dx.doi.org/10.21437/ASVspoof.2024-1.

Wang, Jun-You, Lee, Hung-Yi, Jang, Jyh-Shing Roger, Su, Li, 2023. Zero-shot singing voice synthesis from musical score. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 1–8.

Wang, Yuxuan, Skerry-Ryan, RJ, Stanton, Daisy, Wu, Yonghui, Weiss, Ron J., Jaitly, Navdeep, Yang, Zongheng, Xiao, Ying, Chen, Zhifeng, Bengio, Samy, Le, Quoc, Agiomyrgiannakis, Yannis, Clark, Rob, Saurous, Rif A., 2017. Tacotron: Towards end-to-end speech synthesis. In: Proceedings of Interspeech. p. 4006.

Weiss, Ron J, Skerry-Ryan, RJ, Battenberg, Eric, Mariooryad, Soroosh, Kingma, Diederik P, 2021. Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5679–5683.

Welch, Poppy, Williams, Jennifer, 2024. Privacy considerations for wearable audio-visual AI in hearing aids. In: Proc. AVSEC 2024. pp. 47–48.

Williams, Jennifer, Azim, Tayyaba, Piskopani, Anna-Maria, Chamberlain, Alan, Zhang, Shuo, 2023a. Socio-technical trust for multi-modal hearing assistive technology. In: IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops. pp. 1–5.

Williams, Jennifer, Banchs, Rafael E., Li, Haizhou, 2013. Meaning unit segmentation in english and Chinese: A new approach to discourse phenomena. In: Proceedings of the ACL Workshop on Discourse in Machine Translation. pp. 1–9.

Williams, Randi, Machado, Christian Vázquez, Druga, Stefania, Breazeal, Cynthia, Maes, Pattie, 2018. "My doll says it's ok" a study of children's conformity to a talking doll. In: Proceedings of the 17th ACM Conference on Interaction Design and Children. pp. 625–631.

Williams, Jennifer, Pizzi, Karla, Das, Shuvayanti, Noé, Paul-Gauthier, 2022. New challenges for content privacy in speech and audio. In: Proceedings of Security and Privacy in Speech Communication. pp. 1–6. http://dx.doi.org/10.21437/SPSC.2022-1.

Williams, Jennifer, Pizzi, Karla, Noe, Paul-Gauthier, Das, Sneha, 2023b. Exploratory evaluation of speech content masking. In: 15th ITG Conference on Speech Communication. VDE, pp. 215–219.

Williams, Jennifer, Yazdanpanah, Vahid, Stein, Sebastian, 2023c. Privacy-preserving occupancy estimation. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5.

Williams, Jennifer, Zhao, Yi, Cooper, Erica, Yamagishi, Junichi, 2021. Learning disentangled phone and speaker representations in a semi-supervised VQ-VAE paradigm. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 7053–7057.

Wu, Zhizheng, Kinnunen, Tomi, Evans, Nicholas, Yamagishi, Junichi, Hanilçi, Cemal, Sahidullah, Md., Sizov, Aleksandr, 2015. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In: Proceedings of Interspeech. pp. 2037–2041.

Yamagishi, Junichi, Wang, Xin, Todisco, Massimiliano, Sahidullah, M, Patino, Jose, Nautsch, Andreas, Liu, Xuechen, Lee, Kong Aik, Kinnunen, Tomi, Evans, Nicholas, Delgado, Héctor, 2021. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In: ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge. pp. 47–54. http://dx.doi.org/10.21437/ASVSPOOF.2021-8.

Yan, Ting, Conrad, Frederick G, Tourangeau, Roger, Couper, Mick P, 2011. Should I stay or should I go: The effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. Int. J. Public Opin. Res. 23 (2), 131–147.

Yan, Xue, Yang, Zhen, Wang, Tingting, Guo, Haiyan, 2020. An iterative graph spectral subtraction method for speech enhancement. Speech Commun. 123, 35–42.

Yang, Antoine, Nagrani, Arsha, Seo, Paul Hongsuck, Miech, Antoine, Pont-Tuset, Jordi, Laptev, Ivan, Sivic, Josef, Schmid, Cordelia, 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10714–10726.

Yang, Pinci, Wang, Xin, Duan, Xuguang, Chen, Hong, Hou, Runze, Jin, Cong, Zhu, Wenwu, 2022. AVQA: A dataset for audio-visual question answering on videos. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3480–3491.

Yigitcanlar, Tan, Li, Rita Yi Man, Beeramoole, Prithvi Bhat, Paz, Alexander, 2023. Artificial intelligence in local government services: Public perceptions from Australia and Hong Kong. Gov. Inf. Q. 40 (3), 101833.

Zhang, Chan, Conrad, Frederick, 2014. Speeding in web surveys: The tendency to answer very fast and its association with straightlining. In: Survey Research Methods. 8, (2), pp. 127–135.

Zhang, Guochang, Yu, Libiao, Wang, Chunliang, Wei, Jianqiang, 2022. Multi-scale temporal frequency convolutional network with axial attention for speech enhancement. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 9122–9126.

Zuiderveen Borgesius, F., 2018. Discrimination, artificial intelligence, and algorithmic decision-making. Counc. Eur. Dir. Gen. Democr. 42.