# Conversations in Action — Designing Conversational Agents for Co-Performance

**DISSERTATION**
zur Erlangung des Grades eines Doktors rer. pol. der Fakultät III –
Wirtschaftswissenschaften, Wirtschaftsinformatik und Wirtschaftsrecht
der Universität Siegen

vorgelegt von

Margarita Esau-Held

Tag der Disputation: 22.04.2024

Erstkorrektor: Prof. Dr. Gunnar Stevens
Zweitkorrektor: Prof. Dr. Alexander Boden
Dekan: Prof. Dr. Marc Hassenzahl

# Abstract

Conversational User Interfaces challenge traditional models of human-computer interaction as they shift from visual to auditory communication. While this shift initially seems to simplify human interaction with technology, recent user studies have shown that commercially available Voice Assistants for homes and mobile phones often fail to meet users' expectations as interactive resources and assistants. The first attempts to improve interaction design have focused on emulating human-human conversations and behaviors. However, there is still a lack of design guidelines for transitioning from information design to communication and interaction design. This work aims to address this gap by conducting a formative study and four design case studies to understand users' interactions, perceptions, and expectations to derive implications through grounded design. By viewing human language use as both cognitive and social, this research explores the design space of conversational agents using the theoretical lens of Social Practice Theory. Specifically, conceiving conversational agents as carriers of practices in co-performance with humans contributes to the design of supportive and interactive resources. Moreover, the sonification and multimodality of interaction complement voice-first interactions and thus enrich conversational user experiences and further enable engaging interactions. In summary, this thesis contributes to the following fields:

- For **CUI**, this work uses the theoretical concepts of informative, communicative, and expressive behaviors to analyze and inform the design of conversational agents while providing design implications for future voice-first interactions.
- In line with **Practice-Based Computing**, this thesis presents four design case studies that empirically investigate human practices as performances and entities to derive design implications for conversational agents that act as Carriers of Practices in Co-performance with humans.
- For **Practice-Based Computing**, this research contributes alternative visions for ubiquitous conversational agents that facilitate engaging and meaningful interactions by being proactive and multimodal while matching the home practices and needs of humans.

# Acknowledgements

Research is a great collaborative endeavor, and this work and thesis would not be successful without the following people. I would like to thank Gunnar Stevens and Alexander Boden for their unwavering support of my curiosity by giving me the place and space to follow my research ideas and also by teaching me how to put my thoughts on paper through writing, I have overcome one of my biggest personal challenges. I am also grateful to all my colleagues and fellow researchers along this journey. In particular, I would like to thank Dennis for his invincible optimism and his countless thoughtful and engaging discussions that made me reach over my lengths. Veronika, my invaluable source of design wisdom, who cheered me until the finishing line. Lena, Jenny, Thomas, Gabriela, Lukas, Felix, and Paul for investing their time to let me have time to rest, think, and write by being the best team I could have wished for. Thank you all for the exciting journey and for making hard times fun, which turned the research group into a place where I feel I belong.

Furthermore, I want to express my gratitude to Nico, who believed in me to start this journey, as well as, Ronda for her insightful feedback and constructive support. A great thank you to all my co-authors, especially to Mahla, Johanna, and Marvin, and all the participants and contributors to this research who have shared their thoughts, resources, and experiences with me to enrich my knowledge, work, and writing.

Last but not least, I would like to thank Benny, my husband and relentless personal debater, sharing his ideas and criticism with me while continuously believing in me, keeping me mentally strong and save by helping me to push through in the right moments and to pace myself in others to regain energy and perspective. I thank my mother Irina, my family and all my friends from my heart who, directly or indirectly, kindly gave their support and energy to me to complete this research from start to finish.

## Related Publications

Parts of this thesis have already been published as conference or journal papers.

- Section 5: Margarita Esau, Veronika Krauß, Dennis Lawo, and Gunnar Stevens. 2022. Losing Its Touch: Understanding User Perception of Multimodal Interaction and Smart Assistance. In Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22). Association for Computing Machinery, New York, NY, USA, 1288–1299. https://doi.org/10.1145/3532106.3533455
- Section 6: Esau, M.; Lawo, D.; Castelli, N.; Jakobi, T. and Stevens, G. (2021). Morning Routines between Calm and Engaging: Designing a Smart Mirror. In Proceedings of the 5th International Conference on Computer-Human Interaction Research and Applications - CHIRA; ISBN 978-989-758-538-8; ISSN 2184-3244, SciTePress, pages 58-69. DOI: 10.5220/0010658700003060
- Section 7: Esau, M., Lawo, D., Neifer, T., Boden, A., Stevens G. Trust your guts: fostering embodied knowledge and sustainable practices through voice interaction. Pers Ubiquit Comput 27, 415–434 (2023). https://doi.org/10.1007/s00779-022-01695-9
- Section 8: Johanna Weber, Margarita Esau-Held, Marvin Schiller, Eike Martin Thaden, Dietrich Manstetten, and Gunnar Stevens. 2023. Designing an Interaction Concept for Assisted Cooking in Smart Kitchens: Focus on Human Agency, Proactivity, and Multimodality. In Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23). Association for Computing Machinery, New York, NY, USA, 1128–1144. https://doi.org/10.1145/3563657.3595975
- Section 9: Esau-Held, M., Marsh, A., Krauß, V., Stevens, G. "Foggy sounds like nothing" — enriching the experience of voice assistants with sonic overlays. Pers Ubiquit Comput (2023). https://doi.org/10.1007/s00779-023-01722-3

Moreover, these publications contribute to the presented topic. However, they are not included as section of this thesis.

- Tanja Ertl, Sebastian Taugerbeck, Margarita Esau, Konstantin Aal, Peter Tolmie, and Volker Wulf. 2019. The Social Mile - How (Psychosocial) ICT can Help to Promote Resocialization and to Overcome Prison. Proc. ACM Hum.-Comput. Interact. 3, GROUP, Article 248 (December 2019), 31 pages. https://doi.org/10.1145/3370270
- Lawo, D., Engelbutzeder, P., Esau, M., & Stevens, G. (2020). Networks of Practices: Exploring Design Opportunities for Interconnected Practices. In Proceedings of 18th European Conference on
  Computer-Supported Cooperative Work. European Society for Socially Embedded Technologies (EUSSET).
- Alizadeh, Fatemeh; Esau, Margarita; Stevens, Gunnar; Cassens, Lena (2020): eXplainable AI: Take one Step Back, Move two Steps forward. Mensch und Computer 2020 - Workshopband. DOI: 10.18420/muc2020-ws111-369. Bonn: Gesellschaft für Informatik e.V.. MCI-WS02: UCAI 2020: Workshop on User-Centered Artificial Intelligence. Magdeburg. 6.-9. September 2020
- Jenny Berkholz, Margarita Esau-Held, Alexander Boden, Gunnar Stevens, and Peter Tolmie. 2023. Becoming an Online Wine Taster: An Ethnographic Study on the Digital Mediation of Taste. Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 26 (April 2023), 26 pages. https://doi.org/10.1145/3579459

## Contents

# Part I

# Introduction and Overview

# 1 Introduction

## 1.1 Motivation

*“One cannot not communicate.”*

Paul Watzlawick [344]

Everyday conversations between humans unfold without us even realizing that we are communicating. These days, the growing popularity of commercially available voice assistants like Amazon’s Alexa, Siri, and Google Assist for homes and mobile phones, as well as the everyday use of Large Language Models (LLMs) [19] like ChatGPT pervasively invite us to start conversations or use spoken commands to interact with virtual assistants and technology, in general. Research and corporations use different terms to describe interfaces of this kind of technology, such as Voice User Interfaces (VUIs), Conversational User Interfaces (CUIs), or assistants, such as Voice Assistants (VAs), Intelligent Personal Assistants (IPAs), or Conversational Agents (CAs) [119, 60]. Thereby, the technological advancements in automatic speech recognition, natural language processing, and natural language understanding enable these interfaces and agents to emulate human conversations and communication styles [61, 254, 9]. This 50-year-old vision promises intuitive access and control to digital information and interfaces [263].

Today, users pass on simple tasks like switching on and off lights to VAs by using short prompts. Besides these well-established practices, the adaptation and use for contexts beyond the home, music, and IoT control is stagnant [6, 60], as users have exaggerated expectations regarding the capabilities of conversational agents that do not match with current technological design and promises made by promotions [119, 53, 332]. Due to their human-like design, users perceive conversational agents that employ natural language as intelligent counterparts and conclude that they will easily understand every spoken command or request. However, the current experience of using CAs to support everyday task of more or less complexity, such as cooking, ordering food, taking calls, booking restaurants or appointments, leads to disappointment and abandonment of the technology [53]. Aside from unfulfilled promises to support household practices, insufficient usability leads to misunderstandings and dead-end conversations, e.g., loss of control, limited discoverability of functions, and severe NLP problems [221, 255, 38, 220].

Furthermore, previous studies indicate that copying human-human conversations may even lead to negative perceptions of IPAs [68], e.g., an uncanny valley that triggers fears. Additionally, research shows that attempts to adapt functional or social rules from human conversations to voice interaction design have not been successful, yet [254, 265, 94]. A body of work [61, 254, 207, 68] even questions whether the term *conversation* accurately describes the nature of human-agent interaction, as users tend to perceive and call the relationships transactional and utilitarian. Comparing grounding processes of human-human conversations, users do not sense or discern a mutual and collaborative negotiation of shared knowledge and understanding when talking to CAs [61, 58]. With the conversational interactions limited in length and exchanges [221, 119], there is also a lack of opportunities to develop interpersonal connections and emotional dialogues [53, 246], as shown in work on chatbots like Replika shows [180, 299, 246]. In conclusion, some researchers [61, 254, 207, 68] suggest refraining from mimicking one by one "mutual understanding and common ground, trust, active listenership, and humor as crucial social features in human conversations" [61]. Thus far, some general design guidelines and heuristics for CUIs have been proposed but not fully adapted yet, particularly regarding complex tasks and performances [60, 219, 42, 233, 220]. Additionally, we still lack a thorough understanding of the design space and its parameters [42, 43, 61, 94, 189, 220] that encompasses core theoretical concepts which capture and anticipate conversational user interaction and performances of everyday practices[61, 60].

To match practices and with LLMs on the rise [19], we need to deepen our understanding of how humans expect and imagine future conversations with agents and interfaces beyond the verbal exchange of information. Therefore, the vision of user-centered conversational agents should revolve around voice-first interactions that decouple from habitual screen-based principles and explore the design space of conversations between communicative and expressive behaviors [344, 353] to provide engaging experiences through co-performances with conversational agents. By adopting a practice-based approach [261, 297, 361], we might shift the design focus from basic communication and conversation to creating multi-faceted interactions that equally engage humans and non-human actors. As a result, humans might leverage conversational agents to experience new knowledge, meaning, and competence to progress and enrich their performances of practices.

The design goal of assisting human practices involves users engaging in and learning new approaches. In this regard, the socio-material context is substantial to the situated learning of practices, which points to designing an environment that fosters the negotiation of meaning and knowledge while emphasizing the significance of firsthand experience in performing these practices. To this end, utilizing Social Practice Theory [297, 261] allows for an in-depth understanding and mapping of human practices, specifically analyzing their fundamental elements of meaning, material, and competence. The focus in this work on practices-as-performances and carriers of practices contribute to a thorough understanding of the context that might guide the design efforts for conversational interactions:

In terms of performance, speaking or communicating can be seen as a part of practice-as-performance that complements and enhances actions or doings in practice. With Schatzki's

[281] words, the practice encompasses a nexus of sayings and doings. For example, meaning in communication arises in relation to the interlocutors and unfolds interactively [344, 15]. Communication as an interaction process consists of several timely, ordered exchanges of messages that go beyond the phenomenon of one-way communication from a speaker to an audience [344] because it requires an engaging counterpart. Additionally, it refers to an intentional behavior while informational behavior is undirected [353, 17]. During the conversation, both interlocutors continuously affect the outcome of the dialogue [16, 15]. In addition to cognitively applying language [59], humans also undergo an intricate learning process to understand and adhere to the social norms and rules of communication, [344].

Against this background, a conceptualization of human performances as potential co-performances [177] with conversational agents might reveal how human and non-human actors align their practices to facilitate the reconfiguration and learning of new household practices. Particularly, co-performance suggests that artifacts can be viewed as active participants in human practices, operating alongside humans and exhibiting a certain degree of autonomy in decision-making and action control [177].

Building on prior work in HCI that has primarily focused on parameters of human-human communication comparing to human-agent conversations [254, 265, 94], and the profound usability issues that impede potential interactions from the beginning [221, 222, 60], this thesis aims to shift design efforts from mere communication to interaction with conversational agents that facilitate the performance of practices and provides engaging experiences.

In general, CAs are predominantly prevalent in smart homes as information hubs and controllers of smart appliances [6, 60]. The smart home industry primarily concerns optimizing and streamlining daily routines to increase efficiency and productivity [3]. However, this narrow focus often neglects user needs beyond functional assistance in monitoring data and controlling Internet of Things (IoT) devices [310, 148, 3]. There is a growing concern that this emphasis on automation and convenience may result in individuals becoming passive inhabitants of their homes, missing out on opportunities to engage in joyful and fulfilling activities [133, 211].

When designing interactions between humans and conversational agents, it is important to consider human agency as a substantial factor in determining acceptance and trust in this emerging form of interaction [3, 131], similar to the experience of personal competence levels [131, 128]. While automation can improve efficiency and effectiveness, it is crucial to strike a balance with human needs for autonomy and competence. However, current research tends to prioritize efficiency over meaningful and engaging interactions, and there is a lack of emphasis on empowering humans in these interactions [310, 291, 5, 75, 128]. In this regard, it is necessary to have proactive agents that can negotiate knowledge and provide personalized alternative actions to achieve strong co-performance. Although there is existing research on proactive behavior in autonomous systems [232, 173], it lacks design implications for voice-first approaches and conversational agents.

Additionally, it has been suggested that incorporating multi-modality into voice-first approaches can address some of their limitations and complement the proactive behavior of conversational agents [197, 221, 351]. Further studies indicate that the embodiment of the conversational agent can benefit by establishing emotional connections and bonds with its users [61, 33, 35, 178]. However, to avoid putting emphasis on features that contribute to anthropomorphism, exploring visual and auditory multi-modal interaction styles and information representation might enhance the usability and user experience of the conversational agent. Research [223, 238, 239, 197, 351] has indicated that visual feedback, such as organizing and visualizing information hierarchically, can support the adoption of voice interaction.

Despite the limited attention given to integrating sound in conversational agents, there is potential for the sonification of data to enhance the auditory user experience [49]. Sonification of data, or "the use of non-speech audio to convey information" [201] can provide subjective sensations and allow for individual interpretations while still communicating relevant information [271]. However, to derive practical and successful design implications, we must explore the risk of misinterpretation [271, 152] and the potential for unclear or diffuse information. The design challenge lies in encoding information and meaning to enrich the interlocutor's experience while ensuring accurate decoding. Nonetheless, fostering engaging experiences beyond functionality by inducing emotions, sound, and music are promising narrative design parameters to evolve speech-based interaction [292]. Research in the HCI community [191] calls for scientific methods to ensure consistent and replicable outcomes in this research area.

In particular, with voice and speech being emerging design materials, it is crucial to examine how users perceive the intelligence of conversational agents, what they consequently expect, and how these agents can effectively assist and enhance human practices [333, 332, 119, 53, 289]. Consequently, this thesis is driven by the following research questions:

- **RQ1** How might Voice Assistants become Co-performing Agents next to humans?
- **RQ2** How can we design for a conversational co-performance of practices?
- **RQ3** How can multimodal agents contribute to an engaging co-performance?

## 1.2 Areas of Contribution

This thesis addresses the previously described research gaps by empirically investigating and developing conceptual and theoretical implications to guide the design of conversational agents. Therefore, the presented studies are grounded in the profound work of the Human-Computer Interaction community and contribute to the research fields of Conversational User Interfaces, Practice-Based Computing, and Personal and Ubiquitous Computing.

Research in **CUI** focuses on transitioning from human-human conversations to interactions with conversational agents based on the change of "its perceived norms, rules, and expectations" that are predominantly valid for conversations between humans [61]. In this light, previous work highlights the lack of robust design guidelines for this emerging technology of voice-first interaction [60, 219, 42, 233, 220]. Commercially available systems like Amazon's Alexa or Google Assist present additional challenges due to their multi-component nature [6, 289, 53]. The absence or limited use of visual output channels complicates the adoption of new interaction paradigms and the cognitive processing of information [223, 238, 239, 197, 351]. Therefore, this work contributes to the conceptual design investigation of current system use and agent behavior. User-centered design approaches are employed to uncover expectations for intelligent assistance and anticipate future roles of conversational agents and interactive resources for household practices.

Secondly, this work draws from **Practice-Based Computing** [361] to shift the design of conversations between human and conversational agents to interactions. Practice-based computing proposes to follow a three-step approach called Design Case Studies [362], which builds the design on previous extensive empirical investigations to understand the human context and is followed by evaluations of its consecutive design. In this regard, the development of technology artifacts centers around the actual sayings and doings of humans [281]. Social Practice Theory [262, 297] describes human practices as socially evolved entities that humans perform regularly and vary in their elements of materials, competences, and meanings [297]. So far, conversational agents offer mostly single-command interactions to control IoT, music, and search [6]. By adopting a practice-based approach, conversational agents can evolve to support situated and experience-based learning, knowledge negotiation, and decision support. Applying the notion of co-performance [177], this study extends the capabilities and role of conversational agents by using metaphors of human sensing, thinking, and acting [250]. By conducting four extensive Design Case Studies that investigate home and household practices, this work contributes with implications for design to create opportunities for co-performance and an engaging experience with conversational agents. The comparative analysis of the findings [270] demonstrates how to structure conversations around meaningful performances of practices with relevance to its inhabitants. Further, the evaluations highlight the impact of such systems on the users' competence, autonomy, and personal agency.

The current research in **Personal and Ubiquitous Computing** primarily revolves around the functional implementation of IoT design concepts [310, 148, 3, 310, 291, 5, 75, 128]. These studies often involve automating tasks to enhance smart homes, such as monitoring energy consumption, ensuring home security, and controlling lights and music [6, 60, 310, 148, 3]. However, as the digitalization of workspaces extends to homes, new approaches need to respect the home as a private space for leisure, well-being, and household practices that are mostly not strictly regulated but not the goal of passive consumption either [133, 211]. Other researchers from this field emphasize the need for less technophile users who want to share agency with autonomous devices in their homes when they can engage in practices that they enjoy [133, 211]. This work contributes by exploring the design space while using multimodality and sound as design materials to create engaging interactions. Multimodality eases not only the transition from graphical user interfaces to voice-first systems, but also allows for the expression of proactive behavior by home assistants. The findings demonstrate that multimodality is a suitable design choice for balancing human agency and the proactive behavior of conversational agents. Additionally, integrating sonic elements enhances the user experience, although managing ambiguity presents a challenge.

## 1.3 Structure of the Thesis

Part I of the thesis introduces the research objectives in Chapter 1 and continues with a thematical and theoretical summary of related work in Chapter 2. Further, Chapter 3 describes the research design and methodology of this work.

Part II begins by examining user interactions and expectations towards commercially available voice-first systems in Chapter 5. Next, four design case studies investigate different user contexts for conversational agents. Chapter 6 focuses on enhancing well-being in smart homes, especially multimodal assistance in the bathroom. The second (cf. Chapter 7) and third (cf. Chapter 8) design case studies explore interactive and engaging support for food-related practices. Finally, Chapter 9 complements the work with a design approach to enrich the user experience of conversational agents through sound. All five research papers have been published in peer-reviewed journals and conference proceedings.

Part III summarizes the main findings to derive and discuss concepts and implications in comparing the user contexts and practices examined in the study. Finally, the thesis concludes by discussing the contributions made, acknowledging any limitations, and suggesting potential areas for future research based on the existing literature.

# 2 Related Work

## 2.1 Conversational Agents and Interfaces

As outlined in Chapter 1.1, speech-based interaction is on the rise. However, we need to consider several aspects to speak of effective and engaging human-machine conversations. The following chapters describe the evolution from the use of language to conversational interaction by providing insights into general conceptualizations of assistants and conversations, challenges in actual use and interaction, and expectations related to the assumed intelligent behavior of assistants.

### 2.1.1 From Speech to Conversations

Voice Interaction Design is an advancing field of interaction with the particular concern of conveying information using Conversational Agents (CAs) [119, 206, 60]. In general, we can differentiate CAs by their conversation style, knowledge, functions, and embodiment [119]. The evolution of such agents is an ongoing process that reaches back to the first chatbot that was called ELIZA, developed by Joseph Weizenbaum in 1966 [348]. The Chatbots' disembodiment allows for multiple projections by its users depending on the kind of person reading and interpreting the generated text [119]. Predominantly deployed on commercial websites to offer personalized service or interactive FAQs, they are task-focused with narrow but in-depth domain knowledge, engaging in fast-paced conversations by reducing turn-taking and exchanges to a necessary minimum [119].

In addition, Voice Assistants (VAs), such as Amazon's Alexa, use human voices and languages for input and output of short commands and reading out information [119]. The embedding of VAs in smart homes and standalone devices contributed to the (commercial) success of VAs. Human likeness in embodied agents even increases as they aspire to mimic the physical appearance and behavior of humans the most. For example, they employ human features to communicate, such as voices, gestures, and facial expressions [119]. Lastly, Intelligent Personal Assistants (IPAs) is a general term for assistants that aim to offer personalized responses to queries, concerns, or general information requests, e.g., booking appointments or engaging with further functions and applications installed on the mobile phone [119]. They provide extensive and rather general knowledge to answer multiple kinds of questions but keep the overall conversation brief to non-existent. Notably, IPAs claim to assist users in their daily lives through a broad range of interactive services via speech and actions [119].

Previous studies identified fundamental elements like "mutual understanding and common ground, trust, active listenership, and humor as crucial social features in human conversations" [61]. Humans' ability to engage in conversations builds on a life-long process of learning and adapting not only language but also the social rules of communication [344, 59, 16]. Human-human conversations are believed to be inherently smooth because human speakers anticipate expressing themselves effortlessly and unambiguously. Many advantages are expected here, namely fast exchange and little adaptation to the technical specifications or environment [60, 6].

However, spoken-language interaction challenges the cognitive load and active listenership of users [254, 61, 298, 243]. In contrast to familiar visual interfaces, voice interaction frequently lacks explicit markers to start and progress interaction [271] but makes users believe that they can communicate the usual way as they do with humans. Many design approaches aim to successfully emulate human-human conversations through the use of anthropomorphic effects [9, 68, 333]. Researchers [61, 68, 254, 263] emphasize the potential drawbacks of machines that closely resemble humans, as they raise concerns about negative perceptions and question the replicated extent of human-likeness. Humans so far tend to perceive conversations with machines in a transactional and utilitarian way, currently rejecting the achievement of mutual understanding, of goals, or of relationship building [61, 53, 332].

By studying characterizations of human-human conversations [344, 61, 17, 15] as well as communicative and informative behavior [17, 353], they might help to illuminate the current design materials of human-agent interaction, that are enabled by computer-generated speech through Natural Language Processing (NLP) [227, 313]. Additionally, the functionality to access information from online resources and distinct design of an assistant personality [188] makes it a new genre of interaction [243]. To further enhance informational and experiential engagement, additional design concepts such as narrative [9] and feedback communication strategies [197] can be employed. As researchers [254, 6] call for voice interactions that must “fit into and around conversations” [264, 243]. In this regard, a strict differentiation of “human” and “machine” [60] behavior dimensions might not be vital. Instead, I propose to follow a generous interpretation of humans’ communicative and informational behavior [353, 17] that might lead to fruitful and human-centered prototyping of machine dialogues and conversational agent behavior (cf. Chapter 2.2.3).

### 2.1.2 Trials and Tribulations with Voice Assistants

The text-to-speech abilities of the latest developments of Natural Language Processing (NLP) technologies in the past decade [227, 313] shaped the advancement of voice-first applications significantly. Automatic Speech Recognition, Natural Language Understanding, Natural Language Processing, Speech-to-text, and generally Machine Learning build the foundation that enables the performance of such systems with a relatively fast response time and availability. With Large Language Models on the rise, we can expect further dynamic, human-like conversational interactions [19]. However, all systems rely on stochastic evaluations that contributed to the advancement of speech synthesis [1, 169, 288, 286] and voice design [284]. that now enable a human-like modulation of voices [284], less bothering [73], more endearing [39], more charming [347], or implicitly conveying context [284]. The new possibilities also provide designers the chance to experiment with gender stereotypes [312], facilitate voice branding [162], or overall enrich the voice experience [168]. The phenomenon of anthropomorphism [72, 81, 179] describes the adaptation of human-likeness and emulation of human intelligence, which contributes to the perceived intelligence of conversational user interfaces by its users (cf. Chapter 2.1.3).

Besides the question of how much anthropomorphism is beneficial to the design of Conversational Agents [179], current research focuses on the general design pattern of voice interaction and user acceptance of domestic CAs. For the past few years, families have been through trials and tribulations when using VAs in their homes but ultimately have embraced them as daily companions [206, 254, 20, 60]. Most users primarily utilize VAs for accessing and operating household appliances and internet-based services like streaming music, setting alarms, checking the weather, or searching for specific information [206, 6, 289]. While VAs significantly impact how people consume and engage with information [206, 53, 89], previous research suggests that users tend to interact with only a limited number of functionalities and do not fully integrate IPAs as household members, or establish strong relationships [53, 254, 197].

One of the reasons for users' hesitations, research [53, 60, 289] suggests their high expectations for speech as a natural language modality, which are not only left unfulfilled after initial use experiences but also greatly disappointed [289, 53, 9, 68]. Further studies revealed that poor natural language processing [119, 197, 221] and serious usability issues [60, 218, 220, 289] inhibit long-term adoption [53, 60, 289]. For example, the absence of audio-visual cues makes it more challenging for users to discover new skills or functions [351], and advanced technologies that require several conversational exchanges are error-prone [119, 263]. This phenomenon is frequently caused by insufficient NLP and speech recognition, system malfunctions, misunderstandings, and unsuccessful feedback [289, 221, 68]. Meanwhile, the current interactions heavily rely on trial and error rather than information recall or visual assistance [221]. Finally, the described functions and designs underperform regularly, which results in poor and non-engaging experiences [53].

Accordingly, the communication between humans and VAs is reduced to almost single commands to control music, online search, and smart home devices [6, 289], instead of engaging in effortless and interactive information exchange [53, 89]. Speech is primarily used for giving instructions, often shortened to keyword search in its phrasing to minimize errors [221]. However, it is necessary for IPAs to adhere to certain principles in order to function as supportive agents [53]. To ensure a positive user experience, virtual assistants need reliable usability [60, 221, 68] and research on the elements that contribute to a charming, playful, meaningful, and engaging interaction. For now, due to previous dissatisfying interactions lacking emotional connection, users expect efficient and convenient interactions but desire a vibrant assistant that can express thoughts and emotions and engage in conversation, similar to a companion.

### 2.1.3 Becoming Conversational Agents

Human speech ability leads to high expectations regarding the intelligence of services and the behavior of conversational agents. For example, especially folk theories of children [188, 363, 179, 83, 102] show that they exploit behavioral references, such as talking and listening, to infer cognitive qualities [363]. They utilize biological analogies, such as mechanical causality or imaginative reasoning, to explain behavioral traits and view reciprocity as a sign of psychological ones. Consequently, children are inclined to allocate CA to a continuum between humans and artifacts instead of just a specific category due to its mixed animated and unanimated elements, considering those as multifaceted. Likewise, humans may assess and understand the actual intelligence and behavior by the system's output [102, 274].

In their literature analysis of what intelligent in intelligent user interfaces actually refers to, Völkel et al. differentiate between the terms "human intelligence" and "technological intelligence" while highlighting that there is no single definition of intelligent user interfaces in research so far [333]. Usually, intelligence refers to a human characteristic formerly introduced by Turing's article [323] with the question "Can machines think?" in 1950 and eventually started a discussion about machine intelligence. In addition, human intelligence is described as the ability to think abstractly, learn, answer questions, and adapt to the environment [333, 181]. In comparison, technological intelligence refers to as "human-machine interfaces that aim to improve the efficiency, effectiveness, and naturalness of human-machine interaction by representing, reasoning, and acting on models of the user, domain, task, discourse, and media (e.g. graphics, natural language, gesture)." [335] that can be summarized as an underlying motivation of adaptation and automation [333]. The further results of the mentioned study [333] show that central aspects of intelligent user interfaces are adaptation, automation, and interaction. While adaptation means to change behavior context-wise, automation refers to independent actions grounded in personal intentions like "voice assistants that perform tasks for the user"[333]. However, most work stresses the aim and purpose of IUI to assist and empower users instead of replacing their skills [333].

The positive impact of perceived intelligence on the interaction between humans and CAs highlights the importance for designers to ensure transparent communication regarding the realistic capabilities and services of CAs. Failing to do so may lead to quick disappointment and a sense of being deceived among users [252, 260]. Additionally, Human-Robot Interaction studies [252, 161, 257] show that incorporating too many human-like attributes such as mimicry and gestures increases the level of anthropomorphism and might result in detrimental effects on the credibility of future collaborations [257]. Humans developed high expectations based on robots' facial features like the dimensions of the forehead and chin, and if not fulfilled, it harmed their relationship building with autonomous systems. To avoid misconceptions and harm, humans desire to comprehend robots' mechanisms of getting, processing, and judging information and data beyond anthropomorphism [257, 260].

Researchers choose different descriptions to attribute intelligent behavior due to varying assumptions and goals associated with the intelligence of interfaces and systems compared to agents and assistants [333, 274]. Against this background, humans tend to conceive interfaces, tools, and systems as adaptive and interactive [333, 274], relative to agents, which they characterize as autonomous and assistants additionally as personal. Accordingly, CAs are considered intelligent not only because of their speech capability but also because of their ability to connect and provide access to different autonomous devices and services in the home.

The ongoing comparison of intelligent human-human conversations reinforces the image of “a talking artificial intelligence” [53] despite the interactions being implemented “preconfigured (...) rather than being a process that unfolds interactionally” [254]. According to users, smart is defined as when CAs are able to execute actions on behalf of humans and effectively support the everyday practices of humans, e.g., to solve issues and show alternative paths that humans did not think of by themselves [331].

However, in line with Wahlster and Maybury’s [335] definition of technological intelligence, smart home technology tends to focus on and promote effectiveness and efficiency by automating and optimizing routines [322, 48, 126]. Usually, this involves autonomous decision-making to save energy, increase home security and/or (self-)monitoring as well as promises to increase comfort, leisure, and entertainment [322, 48, 310, 47, 29]. Despite those advances, research [211, 333, 310, 147] uncovered opposite opinions of users that indicate that off-the-shelf products refer to the home as a “technological space” [147]. Further, less technophile smart home users oftentimes experience the changes in their home environment as “small conveniences rather than substantial support for routines” [211]. Finally, studies [211, 75, 90, 5, 268] point out people that fear becoming inactive in tasks that they formerly enjoyed doing because of home automation. In general, the users’ understanding and perception of beneficial smart technology in the home remains unclear [331, 211], and needs, therefore, more investigation and consideration.

Despite IPAs’ promises of personalized home services [60], their prevalent task-focused design can sometimes limit their ability to provide relevant support for household practices [119], as they can only execute a single task instead of supporting a set of interconnected practices and needs. Nonetheless, prominent advertising like Amazon’s Alexa presents IPAs as multimodal and multicomponent ecosystems that promise interactive services of third-party providers to assist users’ practices [4]. The Amazon platform allows users to activate and control those skills and facilitate conversational access to information. Similar to applications like LINE, WeChat, or KakaoTalk [304], this organization and access to services resembles so-called “supper apps” [51, 230, 304] or mega-platforms [51]. Essentially, they act as “a single app [that] takes over an array of other services such that it becomes a platform to support all platforms” [51]. Besides communication as a standard function, it offers various services that range from banking to e-commerce, assisting in different everyday situations. Either app developers themselves or third-party developers introduce new skills by utilizing the app interface, as seen with Didi, the Chinese Uber, in WeChat [103].

Similarly, systems like Amazon Alexa allow third-party developers to create and publish additional services [166], evolving into orchestrators of third-party applications that are, e.g., responsible for making reservations or calls. At first, the advantages seem to be a win-win situation for businesses and consumers: strengthening their economic position by increasing their user base and gaining access to a wealth of data [230], while no need for additional logins, reduced loading times [103], and integration with relevant payment services [304] for users. Unfortunately, the current application design of voice-only interaction and provided skill set misfits user practices and leads to further user frustration [289]. In this light, this thesis aims to investigate how voice assistants may become conversational agents that support humans at home.

## 2.2 Conversational Performances and Practices

So far, design approaches focus either on emulating single tasks or conversations. To understand the challenges and pitfalls in designing conversational agents, I argue to conceive and use conversational agents as an engaging counterpart in social practices [261, 297]. The lens of Social Practice Theory enables to explore the design space by focusing on the doings and sayings in the performance of practices [281]. First, it aids in understanding what user practices are and how they unfold. Secondly, it offers insights into how CAs might contribute to the performance and evolution of user practices, particularly when individuals are engaging in unfamiliar or new ones. Moreover, according to Social Practice Theory, engaging in the performance of practices facilitates transmitting and applying knowledge and skills embedded in those practices. In the following, this chapter outlines the basic concept of social practices as an entity and as a performance.

### 2.2.1 Social Practice Theory

Reckwitz [262] defines social practice as “a routinized type of behavior which consists of several elements, interconnected to one another: forms of bodily activities, and mental activities, ’things’ and their use, background knowledge in the form of understanding, know-how, states of emotion and motivational knowledge”. Building on his work amongst others [262, 281, 30], Shove et al. [297] provided a framework for social practice theory, emphasizing the interplay between competence, meaning, and material elements within practices, especially in the event of practice evolution.

As a result, Social Practice Theory enables the analysis of human practices by examining the three fundamental components of competence, meaning, and materials [297]:

- **Competence** refers to “know-how, background knowledge and understanding” [297]. However, a crucial distinction is that knowing represents practical or deliberate skills, a shared understanding of what a good performance of the practice is [112], and equally, it also means to know how to apply the skills by yourself [297, 342].

- **Materials** are “encompassing objects, infrastructures, tools, hardware and the body itself” [296]. Besides Schatzki’s understanding that “practices are intrinsically connected to and interwoven with objects” [282], further research agreed that physical things and objects are considered as material as well [272, 297].
- **Meaning** describes “the social and symbolic significance of participation at any one moment” [297]. Furthermore, Schatzki [283] emphasizes that practices have a history and a location they originated from or evolved in, contributing to the meaning of each practice.

Building on the elements, the interconnectedness and consistent configuration of the elements within a practice-as-performance contribute to the formation of a practice-as-entity, as described by Reckwitz [261]. While performances may vary slightly, they are still instances of the same practice. Past and present performances constrain and define future reconfiguration of practices [297]. Shove et al. [297] argue that “practices-as-entities are shaped by the total of what practitioners do, by the variously faithful ways in which performances are enacted over time and by the scale and commitment of the cohorts involved.” For example, everyday household practices, cooking are integrated, dispersed, and interconnected [184, 343], requiring a contextual and situational understanding to successfully unpack and apply knowledge [297, 349].

Hence, *cooking as an entity* might be understood as preparing something to eat by combining different materials, competences, and meanings associated with the purpose of cooking. These elements can differ across cultures, and even within families or individuals, there may be unique tools, competences, or meanings that are either unknown or have been established over a long period of time. For example, cooking Kaiserschmarrn is a traditional recipe in Austria but is also prepared in southern Germany. The Austrian version requires preparing the dough in a pan, followed by baking it in the oven which impacts the taste of it, whereas in German homes, it will be served without using the oven. However, when people talk about cooking a Kaiserschmarrn, they will have a fundamental and mutual understanding of the practice. Meanwhile, cooking as an entity might encompass a nexus of further practices-as-entities.

As an instance of cooking, individuals might use specific tools as materials in *cooking as a performance*. For example, they will use a fork to produce gnocchi out of dough to comply to a specific form to be identifiable as gnocchi. This family-based and traditional recipe follows rules and norms that are externally visible. Additionally, there may be only certain sauces considered appropriate to serve with the gnocchi, which is based on their compatibility with the shape of the dough to reinforce the taste of the dish. If a son prepares this dish for the first time on his own, his performance still belongs to the practice of cooking as an entity. It will be widely understood by the community in and outside the family because of the shared meaning between generations.

Following Schatzki’s [281] understanding of practice as a performance, he emphasizes practice as a nexus of saying and doing that is rooted in his broader philosophical work on social practices and the intertwining of language and action. He views language not as a separate or isolated entity but as profoundly embedded within practices. When humans engage in a practice, they use language to communicate, coordinate, and make sense of their actions, contributing to enact and sustain practices [281]. Against this background, understanding practices as performances by CAs can reveal the nexus of communication and the embodied actions of humans, as further elaborated in Chapter 2.2.3.

Unlike traditional HCI approaches that treat interaction as a product of the user, practice theory takes a different perspective by considering the user as shaped and formed by the interaction practice in which they are involved. Subsequently, Shove et al. [297] highlight *carriers of practice* in practice theory that enable society to perform and sustain these practices [297, 30, 183, 262]. According to Reckwitz [262], the carrier is an embodied agent or entity that engages in a particular practice. This carrier can be an individual, a group, an organization, or even a technological artifact, for example, CAs (cf. Chapter 2.2.2). In general, carriers of practices serve as the anchors or embodiment of practices, determining whether practices will persist or undergo change [297, 30, 183]. According to Shove et al. "(...) practices-as-entities are defined by the performances of changing cohorts of carriers. Individuals are constantly taking up and dropping out of different practices as their lives unfold." For practices to endure across generations, new carriers must constantly emerge, starting as newcomers or novices [183] who engage with the community and practice, eventually replacing older carriers [297, 21, 30, 183]. This process of succession and learning can often lead to conflicts, as novices bring new ideas and motivations that may result in the reconfiguration of the practice entities [183, 96, 21].

In summary, Shove et al. [297] characterize social practices by the two prepositions: "The first is that social practices consist of elements that are integrated when practices are enacted. The second is that practices emerge, persist and disappear as links between their defining elements are made and broken.".

### 2.2.2 (Re-)Packing and Applying Knowledge

Carriers of practices are necessary to share and circulate competence and knowledge, including the know-how of knowledge application [262, 297]. This process involves negotiating the meaning, materials, and competence of the practice early on, even before the actual performance takes place. This proactive approach ensures the successful transfer of know-how and the establishment of shared understanding [297]. Shove et al. distinguish between the moving of knowledge and its actual decoding as two separate processes [297]. The movement of knowledge, whether physical or virtual, depends on its manifestation, while decoding refers to the ability to apply that knowledge. The challenge arises as the know-how itself is a product of prior experience and therefore not always and everywhere available [297, 84]. This personal and collective experience is directed toward practitioners who already own some initial competence.

Against this background, further researchers propose a distinction between two types of knowledge: embodied [212, 108] or institutionalized [108]. Embodied knowledge is grounded in subjective experience and the negotiation of the “meaning of words, actions, situations, and material artifacts” [108, 109] with others. In these terms, the “competence-to-act” [108, 109] emerges from the physical performance of actions while experiencing the immediate and ambient sensation and perception of those actions, and the final outcome [212]. On the other hand, institutionalized knowledge refers to theoretical rules, approaches, and formalized regulations that are typically made explicit by authorities or organizations [115, 109, 22, 21]. Further, parts of knowledge can be stored in physical or virtual forms, even when not in use [297]. Cooking is a proper example of a practice based on the intuitive negotiation of embodied knowledge, individual preferences, and formalized instructions [14, 108].

Relative to the endeavor, deliberate learning or study often stresses perfecting a skill through repeated practice and effort [297]. Regardless, Lave and Wenger [183] argue that learning already occurs during the performance of practices, as it is rooted in the circulation of competence [297] and the negotiation of meaning [350]. Hence, carriers’ responsibility to effectively circulate knowledge means to reverse the process of abstracting local knowledge when in transfer to new environments [297]. This reversal involves reconfiguring competence in relation to other elements, such as material and meaning. With competence and meaning being socially and culturally contextualized components, the new practitioner’s enactment of this knowledge requires understanding the previous codification of competence to decode it successfully [297].

This understanding and retention of knowledge occur as part of social relationships enacted within communities that significantly shape identity, meaning, and practice [183]. Lave and Wenger [183] describe this paradigm as situated learning. According to them, learning and the negotiation of meaning manifest within communities of practice (CoP): “Communities of practice are groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly.”. Often, this mutual practice is established and shared without conscious awareness. However, effective learning within a CoP requires at least one carrier of practice who can teach the particularities of the practice to others [297, 185]. Further, the engagement in CoP has far-reaching effects on the identity of the individuals participating [183]: “Painting a picture of the person as a primarily ’cognitive’ entity tends to promote a nonpersonal view of knowledge, skills, tasks, activities, and learning.”.

Instead, Lave and Wenger [183] argue that social practice understands the individual as a person-in-the-world and not isolated in its learning experience as acquisition and assimilation. An example of this can be seen in Becker’s study on marijuana users, where novices transform into experts and change their identities through meaningful interactions and engagement in the practices of central members and social representatives of the community [297, 18]. According to Wenger [350], meaning is the central element of social practices and situated learning: “Meaning - our ability to experience the world and our engagement with it as meaningful - is ultimately what learning is to produce”. The negotiation of meaning and decision-making within CoP involves the exchange of past and new knowledge between current carriers of the practice and novices eager to learn and become community members. This process includes the circulation of knowledge and leads to the reestablishment of existing practices or the creation of new ones. It is a dynamic and interactive process that involves the Socratic negotiation of meaning [350].

Finally, apprenticeship should not be limited to a hierarchical relationship between a master and an apprentice [183]. Instead, situated learning proposes that anyone can contribute to the transmission of knowledge by creating diverse learning opportunities changing in contexts and time [183]. Continuing this paradigm raises the question about the extent to which CAs, as interactive knowledge and know-how resources, may be viewed as active carriers of practices that are able to engage in the negotiation of meaning. Or is their role limited by the assumption that they are simplistic technological tools that perform practices and express knowledge externally? However, by adopting the theoretical framework of Social Practice Theory and situated learning, we can examine the design space of CAs in terms of their ability to co-perform practices and allocate competence without delving into the debate of anthropomorphism.

### 2.2.3 Using Language as Performance

While Schatzki [281] stresses the significance of language in social practice, the practice theory by Reckwitz [262] or Shove [297] do not provide a sufficient explication of the concept of communication. In terms of practices, communication is central to the process of meaning-making [281] as humans share and negotiate meanings, understandings, and interpretations specific to that practice. Furthermore, communication is not solely a cognitive process. Instead, it is deeply rooted in the embodied actions and movements of humans experiencing, interacting with, and engaging with the physical world through their bodies. To approach the design of CAs from a nexus of communication and embodied human experience, the following outlines the fundamental principles of the understanding of communication that underlies this work:

Watzlawick et al. [344] proposed to investigate human communication by the four elements of syntactics, semantics, pragmatics [216], and semiotics [45]. This categorization allows us to uncover the syntactical challenges of information transmission regarding “coding, channels, capacity, noise, redundancy, and other statistical properties of language.” [344]. Further, semantics is primarily concerned with meaning that is significant to senders and receivers of messages and, therefore, to sharing information in the first place [344]. Additionally, researchers [344, 26] describe humans not directly to communicate but as an individual who "engages in or becomes part of communication. He may move or make noises ... but he does not communicate. In a parallel fashion, he may see, he may hear, smell, taste, or feel-but he does not communicate. In other words, he does not originate communication; he participates in it.". Consequently, communication and its meaning unfold interactionally in relation to its interlocutors [344]. Following the pragmatic dimension of communication, Watzlawick et al. [344] suggest that "communication affects behavior" [344]. It refers to and integrates the personal context and behavior of the interlocutors. Therefore, Watzlawick et al. initially concluded that "all behavior is communication." [344].

However, studies of the past 30 years [17, 15, 16], particularly those of the second author Janet Beavin Bavelas, revised the initial hypothesis and suggest the opposite that "Even if 'you cannot not communicate', not all behavior has to be communication." [17]. Subsequently, communication requires at least two interlocutors who deliberately decide to engage with each other, as otherwise, the mere presentation of information not targeted at an audience and without intention does not comprise a communicative behavior. By pointing out the early work of Wiener [353], who studied verbal and non-verbal behavior in 1972, [17] emphasizes the difference of informative and communicative behavior:

- *Communicative behavior*: Behavior with which the person intends to communicate something and to which the other person reacts.
- *Informative behavior*: Behavior by which the person does not intend to communicate something, and that is only interpreted as communication by the person observing the behavior.

In contrast to earlier assumptions [344], negotiating of meaning evolves interactionally moment-by-moment based on calibrating information immediately [16]. The series of actions and responses results in an interactive system that is ordered and determined by the variable of time [344]. The micro-analysis of face-to-face dialogues between humans [15, 16] shows how the occurrence of words, style, facial gestures, gaze, and nodding influence the interlocutors reciprocally and immediately. While the conversation unfolds in interaction, its interlocutors affect the outcome in real-time: "And not just after something has been said, but continuously while the person is speaking. For example, they could show that the way we listen strongly influences what the other person says" [15]. Consequently, people who do not intend to communicate and disengage consciously by showing it may, for example, avoid eye contact or look away [16]. Finally, mutual understanding involves meaning-making between the conversational partners [15], as already work of other researchers [16, 208, 202, 314, 15, 321] indicated.

Notably, communication as an interaction process involves several exchanges of messages going beyond the phenomenon of one-way communication from speaker to listener [344]. Therefore, we should not confuse or equate an audience with a community, irrespective of whether we analyze human-human conversations or design for human-machine conversations. Finally, the relationship between sender and receiver is a far more dynamic process than initially anticipated.

## 2.3 Mutual Engagement through Co-performance

The concept of co-performance adopts the perspective of practice-as-performance to reveal and imply how technology shapes social practices [177]. By embracing this approach, this thesis explores how CAs can be effectively designed as interactive resources to align with human practices. Assigning both stakeholders the competence-to-act, we must pay particular attention to ensure human agency with regard to humans' needs for autonomy and competence.

### 2.3.1 Co-performing Practices

Viewing CAs as co-performers of practices allows us to conceptualize the design space as an enabler for engaging experiences that contribute to the negotiation of meaning, knowledge, and experience. In particular, the perspective of co-performance refers to the paradigm of a beneficial distribution of capabilities and responsibilities between human and non-human actors [177, 110, 111, 163]. Kuijer et al. [177] suggest that artifacts own a certain autonomy in decision-making and control of actions and, hence, may be seen as active performers of practices next to humans. Their study on domestic heating practices examined the evolving role of artifacts from fireplaces to thermostats. In conclusion, smart thermostats retain agency in decision-making over temperature regulation, albeit humans maintain to assess heating adequacy and comfort through their senses and overrule the artifact's decisions [177]. The embodied knowledge and experiences of humans may differ significantly from those of smart artifacts, necessitating active negotiation of practice, meaning, and decision-making processes to achieve satisfactory outcomes that align with user goals such as energy savings and comfort.

To further analyze and specify the design space of conversational co-performance, I will adopt the Sense-Think-Act cycle proposed by Pfeifer & Scheier [250]. The authors provide a framework that allows us to study the mutual co-performance of humans and CAs at three levels. From a machine perspective, intelligent machines need sensors (1) to perceive the environment. Then, they process information by computation (2), and finally, they take situated action (3), which refers to the interaction with and in its immediate surroundings and context [250]. While Pfeifer & Scheier [250] stress the significance of sensors and motors to assign embodied actions to robots, the objective of this thesis is to explore the multi-sensory capturing, thinking, and performing of humans in co-performance with CAs. By focusing on the integral capabilities of the human and non-human actors, we can reveal how this understanding can inform the negotiation of competences and facilitate a human-centered co-performance. Previous studies have shown promising results in using CAs as educational tools or companions [102, 200, 71, 137] but currently lack a comprehensive understanding of how to design effective and human-centered learning environments [137].

Against this background, CAs may act as carriers of practice, facilitating the sharing and transfer of knowledge to humans. While carriers, they lack the physical ability to sense the world around them [250] but consequently think and act as described by the nexus of saying and doing [281]. The concept of co-performance necessitates agency from both CAs and humans in situated decision-making. Kuijer et al. argue that "artificial performers should be considered as a category in their own right and not as (poor) imitations of humans ones." [177]. Consequently, designing CAs requires a profound understanding of users' competences and abilities along the act-think-sense cycle and a design process prioritizizing human agency and experience.

### 2.3.2 Proactive Resources in Balance with Human Agency

Traditionally, CAs have been designed to be reactive to human actions or inputs (cf. Chapter 8.1). Before responding, these agents await users to initiate a conversation or provide specific prompts or queries. Meanwhile, there is a growing emphasis on making agents and artifacts more proactive [60, 90, 157, 53, 57]. Proactivity in conversational agents aims to enhance the user experience by reducing the need for users to initiate every interaction and by providing timely, relevant, and helpful information or actions [173]. This assumption leads us to a further consideration that concerns the balance between automation and human agency. As humans own the autonomy and competence to perform practices, they do not necessarily aim to renounce all actions and control to technology [3] and accordingly strive to sustain space for technological independence [310, 291, 5, 75]. Autonomy refers to one being the cause of actions, while competence means that one is capable of and effective at those actions [131]. Both autonomy and competence are of great value to the individual compared to other psychological needs [131, 128]. Furthermore, "Pleasurable experiences" with technology are fundamentally impacted by users' perceptions of their autonomy and competence [131]. For example, human-food interaction research often focuses on automation and functionality, neglecting the social and individual experiences that are deeply intertwined with food practices [3]. This results in artifacts being attributed with greater autonomy and efficiency rather than empowering humans in their practices through technology while maintaining their agency [3]. Therefore, it is crucial to evaluate proactive behavior and its impact on user experience to guide the human-centered design of systems that align with user needs.

Previous research in ubiquitous computing suggests a shift from proactive computing to proactive people [268]. Engaging artifacts play a crucial role in facilitating meaningful interactions that are inherently motivating, as they provide a collaborative context for decision-making and sense-making through experiential learning [268]. CAs designed as things that "act" [329] follow the approach to create interactive resources that own a social presence and join the tasks of users. The offering of personalized advice and situated support may foster personal and meaningful interactions that contribute to relationship building [151] and establish long-term trust [53]. The aim is to support humans in their meaningful and enjoyable performances of practices situated in the home [268, 133, 310].

In that sense, designing for proactive people does not imply the creation of passive or barely proactive technology. Instead, it demands prioritizing human performances and appropriately balancing the proactivity of CAs. Users associate proactivity, much like speech, with smart and engaging technology. Research suggests [232, 173] proactive behavior refers to the capacity to anticipate and take action before an event occurs and to classify this behavior into different levels such as "none," "notification," "suggestion," and "intervention" [173]. Each level contributes to situated respect for the human performance of practices and requires contextual adaptations, while suggestions and interventions are particularly relevant to advancing human competence. Treating IoT agents as interdependent co-performers in a relationship with humans [57] may lead to an unsettled allocation of control that may "cause a growing tension between human and product agency" [57]. Consequently, the design process must consider the social consequences of the co-performance between humans and technology [57, 177]. In addition to the levels of proactivity, focusing on the capabilities of agents and expressing them through a multi-modal interface can help adapt the agents' proactive intentions to the actions of users without compromising human agency.

## 2.4 Extending Conversational Agents beyond Speech

Co-performance seeks to engage humans in interaction with a proactive conversational agent to explore or master practices. While co-performing, the auditory design guides the sensations and perception of the human. However, there is little knowledge about integrating auditory and multimodal design into existing CUI to enhance the overall user experience. Therefore, the upcoming chapter will address two main topics: the multimodal enrichment of engaging artifacts and agents and the design considerations for creating sonic experiences.

### 2.4.1 Multimodal Interaction

The key to effective communication and mutual understanding is to encode information that enables the recipient to decode the meaning, socially and cognitively [297]. Using language is a familiar and effortless process of encoding and decoding to humans, particularly in spoken communication. Despite the perception of spoken language as a direct means of transmitting information, it still harbors the potential for ambiguity and misinterpretation between humans alike [10]. Unlike written text, the ephemeral nature of speech-based information presents additional challenges, such as increased cognitive load, dead-end conversations, and the potential for abrupt endings when the microphone is closed [60, 254, 298]. The transient manifestation of speech requires listeners to concentrate deeply to process and respond to the information being conveyed [298]. Grice [97] suggests that effective communication practices should consider the appropriate quantity and quality of information, as well as the clear and relevant sharing of information. Adhering to these principles may help mitigate the negative effects associated with speech-based interaction.

Research on auditory interfaces highlights a significant distinction between the cognitive perception of visual and acoustic information [88, 337]. Humans process auditory cues relative to temporal sequences that emphasize the evolution and state of sound and speech over time. On the other hand, visual cues are predominantly understood through spatial positioning in relation to one another, forming a visual hierarchy. Besides, static visual content remains formally unchanged when observed by individuals, but repeated exposure can alter the impression and interpretation of such content [88]. Graphical User Interfaces (GUIs) have been considered the primary interaction paradigm for several decades. Researchers and practitioners have developed and refined consistent patterns, recurring icons, and visual elements that are now universally recognized, ensuring widespread comprehension of content [220, 221].

Consequently, the selection of input and output channels for conveying and presenting information should align with the diverse contexts and cognitive processes of users. Several studies [197, 223, 238, 239, 351] suggest that incorporating visual feedback benefits users' appropriation of CUIs. Further, combinations of interactions enhance the adaptivity of user interfaces. Previous work [197, 221, 351] on combining visual and voice interaction has shown that users appreciate visual cues on screens that illustrate voice interactions or confirm their actions. The opposite, the absence of visual feedback, might even reinforce misunderstandings between CAs and humans. Therefore, integrating voice and graphical user interfaces can enhance the usability and user-friendliness of CAs [238, 239]. Compared to unimodal implementations, this approach increases transparency, flexibility, and efficiency [239, 238]. Furthermore, researchers [279] revealed benefits in terms of system robustness, accessibility, and intuitiveness. However, the goal is to minimize unpleasant and overwhelming interaction efforts. The suggested approach seeks to optimize the advantages of one modality while addressing its limitations by integrating complementary input and output modalities [25, 194, 231, 239].

To prevent user frustration, the contextual adaptation of multimodal interaction might benefit the proactive design of agents while meeting human needs and expectations of the named proactive behavior. For instance, when adopting new practices, multimodality can enhance learning and exploration. Wechsung and Naumann [345] demonstrate that offering multimodal interaction improves the perceived user experience, even if a single modality would be more efficient for completing tasks. However, the currently limited use of multimodality is primarily due to the challenges in development and the lack of experts rather than a lack of purpose [279], such as aligning modalities with the exhibited intelligence of systems [186]. Likewise, the multimodal possibilities of design seldom express the perceived intelligence of IPAs and CAs. Current interface design guidelines for chatbots or similar interfaces that combine speech and graphics could provide valuable insights and best practices [119, 146, 60] to extend the research on multimodal IPAs. Hence, by being multimodal and multi-component, CAs may extend their autonomous capabilities to become engaging counterparts to humans [186, 279, 345].

### 2.4.2 Sonic Experiences

Speech and communication is often reduced to its function to transfer information. However, speech is also a type of expressive performance and storytelling [9] that could enable engaging interactions and experiences.

Experience Design in HCI differentiates between to states, namely, experiencing and a experience [129, 95, 204]. While both states involve emotions, actions, cognition and motivation, experiencing refers to a constant stream of experience, experiencing moment-by-moment which focused on the immediate experience being made [129, 95]. On the other hand, Hassenzahl calls an experience a memory of a lived experience represent by stories that humans remember and eventually alter over time [129]. An experience usually "can be articulated or named: has a beginning and end; inspires behavioral and emotional change" [95]. Further, Forlizzi and Battarbee add the co-experience which extends to creating meaning with others influenced by social presence and interaction. Adapting the interaction-centered view of Forlizzi and Battarbee shifts the focus on "how product interactions unfold and how emotion and experience is evoked." [95]. Therefore, designing for engaging experiences means to create interactive systems that contribute to learning, reflecting and understanding our interactions with our environment and the technology itself [268].

Employing strategies for affective communication [197] can thus further enrich interactions with CAs. Recent discussions have focused mainly on refining NLP and the emerging challenge of cognitive processing of information by humans. However, the Elaboration Likelihood Model [248, 249] highlights that human information processing is complex beyond cognitive and linear processes, differentiated by two routes. The central route involves carefully decoding a message by analyzing its meaning, the persuasiveness of the arguments, and the credibility of the presented facts. On the other hand, the peripheral route involves emotional reactions to the message, where individuals rely on general sensations, peripheral cues, and underlying tones.

The use of speech modulation techniques, such as whispering, adds an additional layer to sonic experiences and helps prevent the perception of CAs as dull or monotonous [243]. Similarly, research in the fields of speech emotion recognition [1, 169, 288], emotional speech synthesis [286] and emotional speech production [187] has investigated the importance of affective and emotional speech in communication [198, 347]. These studies analyze how voices and speaking styles express and disclose emotions such as sadness, joy, anger, tenderness, surprise, and boredom [153, 284, 286]. Correspondingly, speech and voice have an impact on credibility, trust, charisma, attractiveness, likability, and overall perception of personality [284, 290, 347].

However, despite the focus of CAs on communication through auditory channels, there is limited research on how sound can enhance or alter the experience of spoken words. According to Enge [88], sonification represents “the use of non-speech audio to convey information” [201] while visuals involve “the use of computer-supported, interactive, visual representations of abstract data to amplify cognition” [228]. Sonification of data allows for more individual interpretation while still aiming to communicate clear information [271]. The most common design approaches for encoding information via sound in interactions are auditory icons and earcons [27]. However, the design of sonification impacts both the emotional experience and the functionality of the interaction. Beyond representing information, sonification contributes to the inducement of emotions, despite the risk for human subjective misinterpretation [271, 152] and unclear information [271]. Thereby, researchers face the challenge of objectively studying the effects of sonic elements, as they must also consider the emotional and narrative aspects of music and its impact on individuals [292]. Whereas communication is defined by the accurate recognition of information, emotion and intentions [152, 154, 344, 353], concerning music, the psycho-physical relations across musical elements and perception of the listener lead to the expression and experience of emotions [152, 41, 80] while connecting with past experiences induced by the listening to familiar sounds. Moreover, incorporating soundscapes can also have a significant impact, such as enhancing tasting experiences and providing a sense of pleasure [46, 341]. Soundscapes refer to the acoustic environment that individuals perceive, experience, and understand within a specific context [143], serving as notable signifiers to listeners.

Consequently, the design of CAs will need to integrate both the clarity of information and the incorporation of emotions through speech and sound to enrich conversational experiences. The HCI community’s research of sound design [191] emphasizes the need for scientific approaches to ensure reproducible results, as the current field relies heavily on the craftsmanship and artistic skills of sound designers: “Sound design can be described as an inherently complex task, demanding the designer to understand, master and balance technology, human perception, aesthetics and semiotics.” [191]. As a result, sound plays a crucial role in conveying compelling narratives and is an essential component of audiovisual storytelling [276, 56]. Sanchez et al. [49] suggest expanding current conversational design practices “to include more nonverbal and paralinguistic elements” and to emphasize sound as a mode of interaction.

Therefore, we need to develop advanced methodologies and design principles for CAs. Currently, most designs draw from established GUI principles for presenting information without considering the specific aspects of auditory information processing, such as the transient nature of speech, memory, imagination, and user interpretation. Following the call of different researchers [298, 49, 60], this work investigates to take the respective aspects into account to enhance the design of CAs.

# 3 Research Design and Methodology

As outlined in Chapter 1 and 2, the current design of conversational user interfaces and agents lacks comprehensive guidelines and approaches. Additionally, these interfaces are predominantly standalone devices in users' homes. In this regard, users expect a substantial and interactive smart home support for their daily activities. This thesis aims, first, to understand users' perceptions and expectations associated with the smart and proactive behavior of conversational agents, particularly in relation to speech and interactive resources. Secondly, it explores the design space to shift primarily communicative behavior to interactions that provide engaging user experiences. The following three research questions guide the research against this background:

- **RQ1** How might Voice Assistants become Co-performing Agents next to humans?
- **RQ2** How can we design for a conversational co-performance of practices?
- **RQ3** How can multimodal agents contribute to an engaging co-performance?

Therefore, a practice-based design [361] approach was applied to address the identified gaps. The deep understanding and anticipation of practices serve as a foundation for designing artifacts and conversational agents that align with the specific context and user expectations. The particularly open and qualitative approach is beneficial not only to comprehend social practices but also for gaining insights into the usage of emerging technologies that have not yet been tested or adopted by users. Its iterative and participative nature contributes to a design process that prioritizes a diverse set of needs and preferences of users that lead to the identification of design implications.

## 3.1 Research Design

As a result, this work follows a multi-stage process of a Design Case Study **(author?)** [362] that encompasses both ethnographic research activities and design research in the following three stages: (1) Empirical Pre-study, (2) Technology Design, and (3) Evaluation. The initial stage involves in-depth empirical research, such as interviews, observations, or focus groups, to establish a solid foundation for subsequent research and design activities. In the second stage, the insights gained from the research build the preliminary user requirements, which inform the design of the technology. At this point, users are encouraged to be actively involved in the artifact design process, as this should not be the exclusive work of (professional) designers, researchers, and engineers. Finally, a target group evaluates the artifact, ideally in real-world usage scenarios and in the wild. However, these stages are not strictly sequential, and designers may iterate between research and design activities until they are satisfied with the outcomes.

In this regard, the first study served as a comprehensive and foundational prestudy to sensitize and familiarize us with the current use and the emerging technology of voice assistants among the German population. In addition to the overall research design, four of the five studies in this work adhere to the structure of a Design Case Study [362]. In the tradition of Research through Design approaches [104, 106] and Grounded Design [270, 307], this work will provide a comparative analysis of the design case studies. The aim is to synthesize design concepts and implications of all studies based on the different stages and objectives, as discussed in Chapter 10. As follows, Table 1 lists all of our design and research activities by study.

## 3.2 Research and Design Activities

Based on the listed research and design activities (cf. Table 1), the concerned methods and objectives are explained in further detail. The studies have combined and applied one or several methods to obtain prospective outcomes and insights. Moreover, most activities covered research and prototyping purposes as well as for pre-study and evaluation purposes. The subsequent paragraphs introduce these activities from a theoretical standpoint, while the individual case studies of this thesis hold specific information on their implementation.

### 3.2.1 Interviews

Interviews commonly serve as instrument to gather data in the field. They might vary in form, purpose, participants, context, techniques, and procedures [136]. In ethnographic research, interviews are often conducted spontaneously and randomly with individuals observed in the field [301]. These interviews should ideally resemble open conversations while still maintaining a focus on gathering the specific data planned by the researcher. Deep interviews focus on narrow themes and aim for a thorough understanding of the research object, and, hence, follow more often previously created guidelines [155]. Expert interviews present a distinct scenario that empasize the role of interviewees as a representative of a particular group or field rather than their individual viewpoints, such as professionals in the food or health industry [214]. In light of this specialized nature of these interviews, they utilize guidelines as an effective means of control.

In general, interview guidelines support the comparability of data, provide an overview of topics and the interrogation flow, and serve as a comprehensive reference to ensure the covering of intended topics. Additionally, employing open-ended questions fosters a conducive environment where interviewees feel comfortable expressing their personal viewpoints. Unlike closed-ended questions commonly encountered in quantitative surveys, the objective is to prompt participants to provide extensive information, thus enabling the collection of contextualized and insightful data. Therefore, semi-structured guidelines provide a framework for maintaining coherence while allowing deviations to follow the thoughts and narratives

**Table 1**

Details of the conducted research activities.

| Chapter | Title | Research Activities |
|---|---|---|
| 5 | Multimodal Interaction and Smart Assistance | 20 qualitative observation studies in a laboratory setting (paired with thinking aloud and semi-structured interviews) |
| 6 | Designing a Smart Mirror | *Empirical prestudy:*<br>1 focus group involving 5 participants<br>1 diary study involving 10 participants<br>*Technology design:*<br>Contextual analysis and Scenario-based design<br>*Evaluation:*<br>5 heuristic evaluations in a wizard-of-oz setting (paired with semi-structured interviews) |
| 7 | Sustainable Practices through Voice Interaction | *Empirical prestudy:*<br>15 contextual inquiries based on semi-structured interviews<br>6 expert interviews<br>*Technology design:*<br>Iterative role-playing and wizard-of-oz sessions<br>*Evaluation:*<br>15 semi-structured evaluation-interviews based on a video-prototype |
| 8 | Designing an Interaction Concept: Assisted Cooking | *Empirical prestudy:*<br>10 semi-structured pre-interviews<br>10 assisted observations in a smart laboratory setting<br>10 semi-structured post-interviews<br>*Technology design:*<br>Multimodal prototyping based on preliminary design implications<br>*Evaluation:*<br>10 semi-structured pre-interviews (paired with quantitative methods based on standardized questionnaires)<br>10 wizard-of-oz interaction observations in a smart laboratory setting<br>10 semi-structured post-interviews |
| 9 | Enriching Voice Interaction with Sonic Overlays | *Empirical prestudy:*<br>1 user survey with 48 participants (including qualitative descriptions and answers)<br>*Technology design:*<br>Using synthesis of user survey results<br>*Evaluation:*<br>15 semi-structured evaluation-interviews |

of the participants. However, ethnographic studies are particularly interested in the subjective accounts of the individuals as researchers attempt to understand the materialization of attitudes and practices [139].

Additionally, surveys gathered text-based accounts. Depending on the purpose of the data collection, closed questions for quantifying the data and open questions for qualitative responses have been used: In order to capture an unbiased experience [28] such as in the form of a diary study [28], to collect associative descriptions in Chapter 9, or to analyze participants' self-assessments and compare them with later observations of their actions in Chapter 8. However, interviewing remained one of the most common methods in all studies, supplemented by surveys.

### 3.2.2 Focus groups

Focus groups are a dynamic version of semi-structured interviews [155] conducted in a group setting to gather diverse perspectives by fostering discussions among participants [195, 11]. Participants within a group are motivated and inspired to take on and defend their viewpoints. Therefore, researchers may observe the construction of social attitudes. The interviewer's role in a focus group is to facilitate discussions by providing topics and encouraging participants to share their opinions without interrupting the flow of conversation. The outcomes of focus groups may vary depending on the composition of participants. The focus group discussed in Chapter 6 served as an approach to generate user scenarios and create visual representations of technology designs through collaborative brainstorming.

### 3.2.3 Observational Laboratory Study

Compared to interviews, this method focuses on observing and uncovering users' actions, behaviors, and procedures [358, 301]. As mentioned earlier, interview responses can be highly subjective. Sometimes verbal accounts may not accurately reflect participants' actual behaviors and attitudes or may have difficulties to express, for example, embodied knowledge. Therefore, observations serve as a valuable complement to interviews to resemble what participants have said and what they actually do.

Therefore, researchers take field notes while observing participants' everyday routines and interactions with technology [358, 301], and employ additionally the thinking aloud method, which actively involves participants in commenting and reflecting on their actions in the moment of performance. As a result, observations can be categorized as direct or indirect. In direct observation, the researcher is present and may influence the actual performance of the observed practice. This approach enables a detailed field observation in the wild without extra equipment or laboratory infrastructure. Indirect observation limits the possibilities to search for different angles of observation or experience prevalent sensory impressions that influence the performance of practices, for example, when analyzing videos [358, 301].

The studies encompass different variations of observations. While some (cf. Chapter 5, 7, and 8) were preliminary observations to sensitize the design, others were dedicated technology evaluations or design studies (cf. Chapter 6, 7, and 8) that involved Wizard-of-Oz [116, 366] applications. In view of the challenging implementation of the prototypes or artifacts, we conducted some of the studies in laboratory settings.

### 3.2.4 Role-Playing and Wizard-of-Oz

Role-Playing [142] and Wizard-of-Oz [116, 366] facilitates to imitate *intelligent* system behavior. The simple application and adaptability to varying contexts enable testing and observing systems in interaction with potential users. It is a low-cost and rapid prototyping that can be adapted in varying degrees to restrict and extend intelligent behavior or *common-sense* to empathize with the user [38]. Role-playing offers the most freedom of system behavior because the person who is portraying the system does not follow a definite script. For example, in the early design stage of conversational agents, we need first to collect some potential conversation pieces and content to understand dialogues and potential outcomes.

In contrast, further design progress necessitates applying and simulating an amplified set of rules. In this regard, the wizard who makes users believe that the prototype is at a higher functional level than it is will have to start to follow some rules. For example, this might include using a restricted number of intents and answers to provide users or not answering in case of a missing keyword. However, this method allows to iteratively change the system according to users' needs and uncover potential conversational dead-ends or usability issues, like effectiveness and sufficiency of utterances. In conclusion, this method fits a participatory design approach as well as ensures early evaluation cycles to understand users' perceptions and limits. Three of the prototypes presented in this work (cf. Chapter 6, 7, and 8) are based on this approach to different extents.

### 3.2.5 Video Prototyping

Similarly, Video-Prototyping is a widely used technique in the field of Human-Computer Interaction (HCI) for evaluating the conceptual aspects of new artifacts. This approach, as suggested by Diefenbach and Hassenzahl [77], enables the simultaneous observation of various artifact properties, such as functionality, emotions, and relational aspects within the interaction. Additionally, this approach allows for an immersive everyday user experience without distracting participants with usability issues, technological limitations or feasibility, or failed interactions [77, 302]. It is also commonly used to illustrate Human-Agent Interaction scenarios [315, 141]. The purpose of this approach is to facilitate discussions with participants, allowing them to focus on the interaction itself and assess its potential impact and usefulness to them without being distracted by potential usability issues. The objective is to understand how participants perceive and evaluate new concepts and designs, particularly in the case of emerging technologies that undergo extensive design and development processes. Hence, this method is equally suited for early-stage design evaluations as well as conceptual reflections with users.

# Part II

# Exploring Conversations through Design Case Studies

# 4 Introduction

The following chapter of the thesis presents the research studies that have been carried out to fill the existing research gap and answer the derived research questions using the previously mentioned research activities.

In this light, the first study in Chapter 5, which investigated experienced and inexperienced users in their interaction with commercial voice assistants, will respond to the question *RQ1 How might Voice Assistants become Co-performing Agents next to humans?*

While providing insights into multimodal interaction and expectations of behavior and support of the CAs, this study informs all subsequent studies and research questions. Afterward, four design case studies in Chapters 6, 7, 8, and 9 illustrate in detail human practices at home and propose CA prototypes that are different in their interface design and realm of their capabilities. Nonetheless, all of them followed a practice-based design approach to explore the design space of conversational agents between communication and expression.

In line with *RQ2 How can we design for a conversational co-performance of practices?* the leading design objective was to respect human practice in its elements of material, meaning, and competence to provide users engaging experiences through co-performances with conversational agents. Therefore, Chapter 7 and 8 particularly investigate the conversational co-performance of complex food tasks.

Finally, all studies in this thesis deal with *RQ3 How can multimodal agents contribute to an engaging co-performance?* and provide insights and design implications on how designers and researchers might enrich conversational interactions and experiences.

# 5 Losing Its Touch: Understanding User Perception of Multimodal Interaction and Smart Assistance

## Abstract

Intelligent Personal Assistants (IPA) are advertised as reliable companions in the everyday life to simplify household tasks. Due to speech-based usability issues, users struggle to deeply engage with current systems. The capabilities of newer generations of standalone devices are even extended by a display, also to address some weaknesses like memorizing auditive information. So far, it is unclear how the potential of a multimodal experience is realized by designers and appropriated by users. Therefore, we observed 20 participants in a controlled setting, planning a dinner with the help of an audio-visual-based IPA, namely Alexa Echo Show. Our study reveals ambiguous mental models of perceived and experienced device capabilities, leading to confusion. Meanwhile, the additional visual output channel could not counterbalance the weaknesses of voice interaction. Finally, we aim to illustrate users' conceptual understandings of IPAs and provide implications to rethink audiovisual output for voice-first standalone devices.

## 5.1 Introduction

Intelligent Personal Assistants (IPA) promise interactive services to support users in their everyday life. As their popularity is growing, so are the capabilities as extended services by 3$^{rd}$-party providers [4]. Similar to supper apps [51, 230, 304], IPAs are evolving from a set of main functions, e.g. taking calls or booking restaurants, to orchestrating 3$^{rd}$-party applications, so-called skills (Amazon's Alexa) or actions (Google's Assistant). Alexa is an IPA embedded as a Voice User Interface (VUI) in a standalone device, offering a scope of functions to many households worldwide. However, prior work [6, 289] indicates that users interact with just a small amount of functions and refrain from adopting IPAs as household companions [53].

A great body of research suggests that non-adoption by users derives oftentimes from high expectations of speech as a natural language modality [53, 289, 60]. New interaction modalities and their accorded platforms and ecosystems bring conceptual as well as interactional challenges, as so far none to just a few heuristics exist [220, 174, 298, 60]. Previous research on IPAs and Conversational User Interfaces (CUI) [60, 218, 220, 289] discovered several issues that impede long-term adoption [53, 289, 60], as they are failing natural language processing [197, 221, 119] and a severe lack of usability, e.g. loss of control, missing feedback, limited discoverability, amongst others [221, 255, 38, 220]. Some researchers [223, 238, 239, 197, 351] indicate that visual feedback might support the appropriation of VUIs by users and multimodality, in general, may improve interaction and its adaptiveness to users.

**Research Gap:** So far, research did not investigate the conceptual challenges of IPAs as orchestrators of several components — Skills and the platform itself, as well as visual and auditive user interface components. Besides the work on mobile IPAs and chatbots [119, 146, 60], it is unclear how the potential of a multimodal and multi-componental experience is realized by designers and appropriated by users, so far. This leads us to our two main research questions:

**RQ1:** What are users' expectations and mental models towards the orchestration of skills regarding its operability and functionality?

**RQ2:** How does a multimodal interface design support users' multi-componental experience and interaction?

Therefore, we observed 20 participants in a laboratory setting, planning a dinner with the help of an audio-visual-based IPA, namely Alexa Echo Show. In particular, we selected five 3$^{rd}$-party skills to explore the interaction and perception of potential users in scenario-based tasks that included browsing promoted goods, creating and managing shopping lists, and browsing and accessing recipes. Each of the participants completed ten tasks in total. Based on our observations, pre-, after- and in between interviews, we were able to derive insights about the understanding and expectations of IPAs by breaking down the conceptual design into its key components.

Our study reveals ambiguous mental models of expected and experienced interactions with IPAs and disappointment regarding the current value of multimodal, standalone devices. So far, participants held Alexa as an IPA responsible to maintain communication across all provided functions, including 3$^{rd}$-party skills. We observed switching of skills as a major source of challenges. Instead of counterbalancing speech-based weaknesses, visual and vocal information seem to compete for the participants' attention and reinforced several interaction issues. While the display was appreciated as an additional property, it should be properly considered within an interaction.

With our empirical investigation, we provide insights into how users understand and perceive IPAs managing and operating several skills to provide value as task-oriented support in everyday household tasks. Moreover, by applying the analytical lens of super apps to IPAs, we contribute by showing the importance of considering the several conceptual and interactional components and their relations to each other for the design of multimodal interaction. Based on this, we provide four implications for research and design that reflect this new perspective to develop interactions based on users' practices and their mental models of IPAs as orchestrators of audio-visual modalities and several skills.

## 5.2 Related Work

### 5.2.1 Users' Expectation and Frustration With IPAs

In recent years, Siri, Alexa, and the Google Assistant have grown in popularity and prevalence and support different contexts by offering a broad range of skills (applications) or actions aggregated in a single standalone device. Grudin and Jacques [119] differentiate between virtual companions with a broad knowledge base and engaging conversations, intelligent assistants that can answer almost any question but keep conversations brief, and task-focused chatbots that are narrow but deeper in their knowledge domain and aim at fast and brief conversations. By definition, Amazon's Alexa is an Intelligent Personal Assistant (IPA) with extended capabilities, so-called skills, through partnerships with 3$^{rd}$-party services. Consequently, Amazon provides a platform with Alexa as an assistant, which controls the skills users choose to activate, and is responsible for conversational access to information.

This is quite similar to what Nierborg and Helmond [230] call super apps, or Chen [51] refers to as mega-platforms. Examples for those apps are LINE, WeChat, or KakaoTalk [304]. They are characterized as "a single app [that] takes over an array of other services such that it becomes a platform to support all platforms" [304]. While the base function is communication, additional services range from banking to e-commerce, thus offering support in several everyday situations. Services can be provided by the super app developers themselves but also based on their app interface that allows 3$^{rd}$-parties to develop new features within the ecosystem, e.g., Didi, the Chinese Uber, in WeChat [103]. Similarly, systems, such as the Amazon Alexa, allow 3$^{rd}$-party developers to design and publish additional services [166]. The advantages, at first sight, are clear, providers of super apps strengthen their economic position by increasing their user numbers and having access to a large amount of data [230]. For consumers, on the other hand, there is no need for additional login, less loading time [103], and integration with the relevant payment services [304].

Despite their variety in skills, however, only a subset of Alexa or Google Home functions is utilized and integrated into a user's daily routine because the IPAs often fail to meet high initial expectations [53, 197] — on the one hand, perceived intelligence, and on the other hand, anticipated usefulness. Language reinforces the image of "a talking artificial intelligence" [53] despite the interactions being implemented "pre-configured (...) rather than being a process that unfolds interactionally" [254]. Similar to chatbots, IPAs are designed very task-oriented [119] and are therefore limited in their routine support. This specificity clashes with inflated hopes built up by the use of natural language as an interface typical for human-human conversation [197]. As a result, users often blame themselves for interaction-based errors, assuming advanced capabilities in Natural Language Processing (NLP) of IPAs [197, 221, 119, 68].

However, voice-only interaction does not fit every task performed in a household scenario [289], which further adds to potential user frustration. For example, users have difficulties discovering functions or skills [351] due to the lack of audio-visual affordances, and advanced features often require more conversational exchanges, which are more likely error-prone [120]. Therefore, music, search, and Internet of Things remain as the main command functions in use [289, 6]. However, for IPAs to be perceived as auxiliary entities, they need to be designed accordingly [53]. Therefore, in the next section, we detail the currently known challenges regarding interacting with IPAs.

### 5.2.2 Facing Challenges in Interactions with IPAs Through Multimodality

Most empirical findings of previous speech-based interface studies point out distinct usability issues that lead to frustration and limited use or even non-use [53, 61, 197, 289]. For example, previous bad experiences with Voice User Interfaces (VUI), like missing or wrong feedback or a lack of the transparency of the system [197], impact future use of VUIs. Corbett et al. [66] describe this as a "negative transfer". Miscommunication in VUIs is one of the main problems [221], especially the failure of NLP to interpret user commands correctly. Those issues evoke various workarounds and tactics by the users, such as “Settling, Restarting, Frustration Attempts, and [even] Quitting” [221], trial and error by hyper articulation [255] or keywords [221, 255]. Other frequently discussed issues include “Mapping”, “Visibility/Feedback”, and “Control and Freedom” [220, 38, 193].

In their study, Myers et al. [221] observed users’ search for visual cues on screens in complex, yet stumbling, voice interactions. Additionally, the lack of visible feedback reinforces misinterpretations on both sides. Users often remaining insecure, because they are used to relying on visual confirmation of their action [197, 223, 351]. It is, therefore, reasonable to assume that the combination of visual and speech-based interaction can lead to an improvement in the usability of IPAs: Oviatt [239] describes the advantages of multimodal applications as a general "improvement in usability" [238] and an increase in "transparency, flexibility, and efficiency" [238] compared to unimodal implementations. However, for multiple modalities to be beneficially combined, a robust set of design aides is needed [60, 298]. Yet, post-wimp (windows, icons, mouse, pointer) interfaces like VUI [220, 60, 298] or interfaces for Mixed Realities [174] still lack a uniform and acknowledged set of such design guides, in comparison to traditional GUI-based systems.

After all, research in auditory interfaces points to the main difference in perception of visual and sonic information: visual marks are perceived in space and auditory marks in time [88, 337]. The latter emphasizes the sequences of sound and speech respectively sound waves over time. In comparison, visual content does not change in perception by looking at it for some time. Of course, the meaning behind what is shown might change, but not the content at present [88]. In general, visualization has the advantage to provide a hierarchy of informational content. Additionally, over time research and practice evolved recurring patterns, icons and building blocks to ensure mutual recognition of the content. Conceptually speaking, by using auditive and visual information, we have to ground the use of modalities on its benefits to effectively combine both, and meanwhile counterbalancing weaknesses by complementing both input and output modalities [25, 194, 231, 239].

While there is limited work specifically examining IPAs with a display, related conversational interfaces such as chatbots [119, 146, 60] or other speech-based systems could provide guidance on applicable good design practices: Compared to smartphones, wearables and traditional computers, the most striking difference between voice-controlled stationary devices with a display is that physical attachment is no longer mandatory. Following the voice-first paradigm, IPAs with attached displays seem to be a new class of devices on their own.

## 5.3 Study Design

In this paper, we want to investigate the role of IPAs to orchestrate and operate several skills as platform managers. Further, we want to explore how well speech and visual representation are currently balanced and whether it leverages the overall interaction with the *Amazon Echo Show* by providing visual aids. Therefore, we decided to conduct an observatory study in a laboratory setting. We decided to observe the interaction with household food shopping skills because the category food&drink was amongst the most popular ones in 2019 in Germany [4]. Our test scenarios aimed to investigate a subset of Alexa Echo Show skills provided by $3^{rd}$-party services that were created to support household practices in the context of cooking, such as browsing promoted goods, creating and managing shopping lists, and browsing and accessing recipes. Our goal was to compare the approaches and usability issues surfacing when operating different skills in three use cases: (1) searching for promotions, (2) browsing and finding cooking recipes, (3) creating and (4) editing shopping lists. Participants were asked to solve each scenario with at least two different skills to be able to compare approaches and the implementation of skills across similar tasks. We will further detail the tasks in Section 5.3.2. Overall, each participant completed ten individual tasks during our study.

**Table 2**
Overview of the skill's features we applied in our study.

| Skill name | Features | | |
|---|---|---|---|
| | Promotions | Recipes | Shopping lists |
| REWE | x | x | x |
| real;- | x | | |
| Chefkoch | | x | |
| Kitchen Stories | | x | |
| Bring! | | | x |
| Alexa | | | x |

The investigated skills were provided by retail brands *REWE* and *real;-*, which are well-known in Germany, for browsing current promotions. Additionally, we included skills from *Chefkoch* and *Kitchen Stories*, as they are popular kitchen smartphone apps for accessing recipes. Both services are also promoted as partners by Amazon. We also included the recipe browsing feature of *REWE*'s Alexa skill. Finally, for organizing shopping lists, we used the skill *Bring!*, which is also a smartphone shopping list app, the corresponding shopping list feature of *REWE*'s application, and Amazon's own feature. Table 2 presents an overview of the skill's features we used in our study.

The study was conducted on an Alexa Echo Show 2$^{nd}$ Gen. with a 10-inch Touchscreen and Dolby-Soundsystem due to its popularity and market share [308].

### 5.3.1 Recruitment and Participants

**Table 3**
Study participants (n=20) representing the German demographic distribution for people with an age > 18 years.

| # | Age | Gender | Occupation | Exp. IPA | Alexa at home |
|---|---|---|---|---|---|
| P1 | 30 | f | Service | Yes | Yes |
| P2 | 55 | f | Health | No | - |
| P3 | 54 | f | Law | Yes | No |
| P4 | 41 | m | Finance | Yes | No |
| P5 | 25 | m | Finance | Yes | Yes |
| P6 | 34 | m | Textile | Yes | Yes |
| P7 | 57 | f | Service | No | - |
| P8 | 64 | f | Retiree | No | No |
| P9 | 32 | f | Communication | No | Yes |
| P10 | 33 | m | Social Services | No | - |
| P11 | 22 | m | Apprentice | Yes | Yes |
| P12 | 49 | m | Public Administration | No | - |
| P13 | 58 | f | Housewife | No | - |
| P14 | 45 | m | Health | No | - |
| P15 | 53 | f | Social Services | Yes | No |
| P16 | 44 | f | Logistics | Yes | Yes |
| P17 | 49 | m | Public Administration | Yes | Yes |
| P18 | 40 | m | Communication | Yes | Yes |
| P19 | 45 | f | IT | No | - |
| P20 | 62 | f | Retiree | No | - |

Our participants were recruited with the help of a professional user experience agency. Participants were selected from a pool of 109 individuals based on the following preconditions: Resembling the demographic distribution of Germany in 2019 regarding gender, age (over the legal age of 18), education, income, immigration background, household, and the family type as well as an even distribution of previous knowledge in interacting with voice assistants (having/not having previous knowledge). Table 3 provides an overview of the corresponding data regarding age, gender, current occupation, experience with voice assistants, and if the participants own or use Alexa as an IPA at home. In total, 20 participants (gender: 10f, 10m; mean age: 45 years, age groups: 18-24 (1), 25-34 (5), 35-44 (3), 45-54 (6), 55-64(5)) participated in our study and were compensated with 50 € each.

### 5.3.2 Study Procedure

Our qualitative study was conducted in a laboratory setting in Germany, as depicted in Figure 1. Each session lasted for 60 minutes on average and was both video and audio recorded. The interviewer stayed in the same room with the participant and was communicating via a tablet with a remote observer who operated the recording equipment. During task solving, our participants were asked to verbalize their ideas and concepts using the think-aloud method. We executed the following procedure: After signing a consent form, we introduced our participants to the general study procedure and let them ask questions. We then proceeded with a short structured interview consisting of four questions to learn more about the participants' household practices, previous experiences, and our participants' anticipation of Alexa and its skills to understand current mental models of use and interaction.

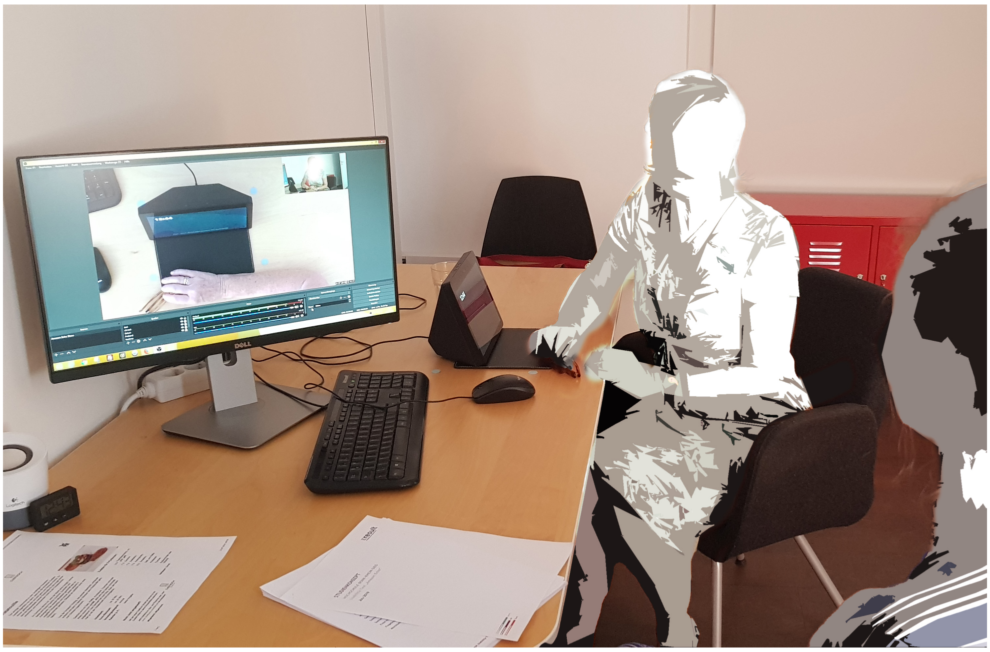


**Figure 1**
User study setting.

After a slot not shorter than three minutes in which participants were asked to familiarize with the IPA, the ten tasks were processed by the interviewer and the participant one after the other as follows:

- Explore current promotions with the skill from (1) *REWE* and (2) *real*,
- Find and browse recipes using (3) the *REWE* skill and (4) *Chefkoch*,
- Create a shopping list based on a printed-out recipe in (5) *Amazon*, (6) *REWE*, and (7) *Bring!*,

- Access, edit and purchase the created shopping list's content with (8) *Amazon*, (9) *REWE*, and (10) *Bring!*.

We adapted and changed the task sequence, but ensured that each skill was tested to allow our participants to solve the tasks as close as possible to their usual way of approaching such scenarios. After each completed task, we asked the participants how they coped with a task, how they felt about the dialogue, whether there were any difficulties, how they rated the comprehensibility of the announcement texts, how they felt about the speaker's voice, how well the application guided the participants, to what extent the content met the participants' expectations, and whether the on-screen display was helpful. Finally, the participants answered closing questions to record what they specifically liked, how their expectations of the Alexa Echo Show were fulfilled, where they identified issues such as potential enhancements, how they would characterize Alexa using a single adjective, and how they would integrate Alexa Echo Show in their daily routine.

### 5.3.3 Data Analysis

For analyzing our data, we used a qualitative approach and conducted a thematic analysis to uncover and present issues in nuance and detail [32]. Therefore, the interviews were transcribed verbatim and coded inductively and independently by two researchers in MaxQDA. Since our interviews were conducted in German, we translated the participants' quotes in this work to reflect as close as possible what they externalized when interacting with Alexa Echo Show.

In the next step, two interaction researchers analyzed transcripts of the voice recordings and the related videos, comparing the commands and visualizations. We focused in particular on obstacles and display use, as well as differentiating between the interactions themselves and the subsequent feedback of the user. Following an iterative coding approach, we finally discussed and grouped codes that emerged from issues that concern prior device expectations, skill activation and functionality, as well as visual and vocal interaction, to analyze the various causes of failure.

## 5.4 Findings

### 5.4.1 Understanding and Interacting With Alexa Echo Show

**5.4.1.1 Expected Range of Features** After the pre-interview, we introduced each participant to the device and discussed their expectations and understanding about Alexa. In general, all participants had quite realistic assumptions about Alexa. The users either owning an Alexa or with prior experience concluded that the main functions like playing music, answering simple questions, and smart home control, work quite reliable. Inexperienced users had similar ideas based on advertisements or stories of friends and family members, but expressed slightly higher expectations regarding the support of activities like online shopping or organizing the household. P9, for example, expected the assistant to directly order the listed items on the device. Others, however, voiced their concern to be "overloaded by advertisement and product placement" (P6) and non-requested orders from Amazon but expected intelligent services:

> “It definitely may not make any mistakes, because if there are suddenly products in my apartment that I didn’t want to order. From portion sizes to whatever else, it has to spoon-feed practically everything to me, preferably repeat it. It has to interact with me and be able to learn. So if I’ve already ordered the things in exactly the same size, then maybe it won’t ask me again. That would be my dream world.” - P17

Yet, most participants were curious about the potential of having more advanced features that would exceed the widely available voice-only devices such as Echo Dot, Siri or Google Home, and implied that functionalities such as video-conferencing, checking the news or watching movies will be possible. Several participants associated the physical properties of the device with a "dismountable tablet equipped with speakers" (P14), or invoked the association with an Amazon tablet or a portable device:

> “So for now, it looks like a classic tablet in terms of design. I don’t know yet. So touching it right now is probably not thought of yet. The back is somehow a bit longer than a classic tablet. I have not seen it yet, I’ll be honest. Well, I’ve seen something similar from Google, but I would imagine that I could take off the monitor at home and actually use it to operate everything that might happen somewhere else on the Internet” - P14

**5.4.1.2 Exploring Alexa** In the exploration phase in which participants were asked to investigate and explore the device, many participants tried to find settings and more information about the scope of functions via touch due to the device’s tablet characteristics. Others started right away talking to the IPA and either used the displayed proactive tips to lead on interaction, e.g. “ask about your favorite singer” (P2), or asked random questions. Overall, the tutorials offered by the manufacturer were considered helpful and well implemented, although not everyone recognized the visually represented tips. Furthermore, participants who were unfamiliar with Alexa had no issues with voice and learned to express simple commands or questions, at least until the end of our session. All participants were expecting to effortlessly browse information online and to get complementary auditive and visual information in a structured manner. The capability to search for and find any information in a short time raised high expectations for some users. Some even were convinced that “there is nothing, she does not know. (...) An almighty device.” (P1).

**5.4.1.3 Alexa as an Orchestrator of Skills** The biggest challenges and misconceptions occurred when participants started to activate and explore 3$^{rd}$-party skills, which expand Alexa's scope of functions. Until that point in our study, participants had not distinguished between the Assistant as a platform manager, the skills as additional 3$^{rd}$-party applications, and the VUI as the main control. However, Alexa functions as the orchestrating service to open, close, and guide a user through installed skills, using the modalities of GUI, too. This concept of changing hierarchies was invisible to our participants and made them insecure about when and which skills were activated and how to leap from one skill to another or go back to the top level. For example, P9 had trouble in understanding how the skills from REWE, Kitchen Stories, and real;- are related and frequently lost the context when they switched between the skills:

> "So, right now I associate Kitchen Stories only with REWE. Well, I have the feeling that I only get to see Kitchen Stories when I use the REWE skill. Hm, it would bother me if I would search something in real;- but it would be shown in Kitchen Stories. (...) Does Kitchen Stories only belong to the real;- or REWE app?" - P9

**5.4.1.4 Ambiguity of Commands** In case of misunderstandings or dead-end conversations, many participants blamed themselves for miscommunication. However, when issues arose, Alexa was identified as being the cause of the issue rather than 3$^{rd}$-party skills. One main problem that led to abruptly terminated conversations or perceived loss of control, were inconsistencies between global and skill-specific local commands. Currently, standardized global commands, like "skip", "next", and "back", for operating the main functions of Alexa and the installed skills are not established. However, learning additional specific local skill commands leads equally to required effort and confusion. For example, P15 got frustrated when commands they previously applied for the same intended action did not work as expected because they were operating a different skill. P18 suggested to keep speech-control local in the currently operating skill until leaving:

> "So, if you are in such a recipe area and ask rather simple things like ‚Back' or ‚Forward', this should also remain in this recipe area and not generally in the system. Not jumping back and forth in the system in general. That would be quite good." - P18

Also, P8 asked for help for the Chefkoch skill to which Alexa replied: "If you wish to pause reading, say, 'Alexa, stop'." However, when P8 followed this advice for the next step of browsing through the recipe and asked Alexa to stop, instead, the whole skill was closed without asking for a confirmation. Participants criticized this behavior and wished for Alexa to confirm such actions before they are executed:

> "So there must be a transition. That I'll get a warning when I'm all the way

> out. (...) So I would like to see the following: 'Do you really, want me to quit the Chefkoch program now?' That I'd get another warning, and then I say, 'no, please stay in. I'm looking for a roast recipe for lamb'." - P8

### 5.4.2 Interacting With Skills

#### 5.4.2.1 Activating Skills

**Initial Challenges Due to Wording** Despite having a built-in tutorial for operating Alexa, the majority of inexperienced users needed the interviewer's support to open their first skill. In particular, due to initial use, we observed that participants had to familiarize themselves with the meaning behind 'skills' first. One issue was Alexa sometimes using very specific interaction wording that creates barriers, especially if users are not native English speakers. As already mentioned in Section 5.3, our study was conducted in Germany and therefore not all participants were fluent in English. However, to activate a skill, you have to know what the English word *skill* means. Symbolically, this stands for an application, but P13 wondered about the meaning until the end:

> "But then she also always says something about 'skill'. And I don't know. She probably can't say everything, either." - P13

Also, sometimes wording did not convey the functionality and, in case no additional explanation was found, participants felt left disoriented. For example, P17 wondered how a shopping list on Alexa Echo Show might behave and what to expect when further interacting with it.

> "The word 'shopping list' is already a specialized vocabulary you have to know. Is it persistent like an online wish list from Amazon, or does it work more like a [virtual] shopping chart [containing all goods] I currently try to order? Will [my goods] be delivered, or do I have to go buy them somewhere? My questions were never answered in the skills, but probably I just could not find it." - P17

**Challenges Due to a Mismatch of Naming** Additionally, participants had to learn the different activation phrases or names for the respective skills which did not align with previously learned conventions such as brand or service names of their everyday supermarkets like REWE or real;- or digital services like Chefkoch or Bring!. Frequent issues occurred in case of misspelling skill names because of slight differences in naming: Some skills are established brands or service providers in Germany and well-known for their websites or smartphone applications. Participants are used to saying "chefkoch.de", "real.de" or frequently also "REWE app". However, only Chefkoch is activated by "Chefkoch.de". This led to participants failing to open skills they required because they either assumed that naming the brand should be enough or they handled it like a web search by spelling out the web address:

> P1: “Alexa, open REWE. W-W-W Dot REWE Dot D-E.”
> Alexa: *dismissing sound*
> P1: “That’s why I don’t like it. It takes far too long.”

**Activating Unexpected Skills** Previous work already reported on common NLP issues and misunderstandings [6, 289, 221], which we could equally observe as the main source for activating unexpected skills amongst others. Hence, we will just point out issues that add to the existing body of knowledge. Sometimes, participants encountered situations in which they could not understand why a certain skill was activated. Such a situation was frequently when another keyword was contained in the original vocalized phrase, such as "Alexa, show me a recipe for Ratatouille in REWE" (P5). Due to the existence of competing skills, Alexa opened an unexpected skill related to ’recipes’. In P5’s case, Alexa opened REWE only after two times redirecting P5 to Kitchen Stories without offering help to recover from that error:

> “Eh, now I definitely did not end up with REWE. Kitchen Stories? That’s somehow something else then.” - P5

Often, the participants associated a brand or service’s name with quality content matching their needs and felt patronized if Alexa directed them to non-requested skills. As it was frequently voiced, users wanted their decision to be respected and executed. While some participants stayed relaxed as long as they received information matching their current context and intended interaction, e.g. when selecting recipes, the majority of our participants were irritated, such as encountered by P13 when they attempted to get information about a recipe:

> P13: “Alexa, What ingredients do I need for chili carne?”
> Alexa: “Aroma Designer is enabled. Aroma Designer may contain content appropriate for adults only. Do you want to open it?”

#### 5.4.2.2 Interoperability and Functional Scope of Skills

**Associated Context of Use and a Skill’s Scope Caused by Previous Brand Experiences** We observed several times that participants expected a skill to be consistent with the brand experience they knew from other platforms. For example, users expect similar content in the skill as on the conventional online platform like Chefkoch and were frustrated if this was not the case (e.g. P18). Further, participants also expected skills to have the same or at least a similar structure, further informational content, and scope of functions compared to the original website or respective smartphone application. However, these expectations were often dashed, for example, due to underlying data models that were not compliant with voice interaction. Chefkoch, for example, is based on user-generated content that is created and shared on its website. In comparison to Kitchen Stories, which is a recipe and food app with mainly curated content, Chefkoch has less structured recipes and databases that cannot be easily accessed on a platform like Alexa without data cleaning and unification of the recipes. In contrast, Kitchen Stories provide in its app detailed step-by-step recipe with matching ingredient lists for every step, as well as in most cases short videos for the preparation of food. This leads to frustration amongst participants who cannot find or access recipes they know from Chefkoch.

**Expected Skill Scope Due to Daily Routines** At the beginning of our study, our participants already mentioned situations of their daily routines that would be interesting to be supported, such as creating shopping lists for their preferred grocery store or supporting with cooking via describing single steps. With the activation of the REWE skill, participants started to think about their online and offline food purchasing practices from either encounter in grocery stores or previous orders through REWE's online shops. For example, some participants started to describe the organization of the supermarket or the online shopping functions from the website. If these ideas did not match the implementation of the skills, participants were most likely disappointed. P8 imagines his ideal process based on what he knows from online shopping and browsing special offer leaflets at REWE:

> "Especially when I go to the supermarkets that I get short responses. Including pictures, of course. And if I don't want to see that, like the cheese or the pudding, I then say ‚Continue', and that it is subdivided like a menu. Now, if we stick to dairy products, I know that at number four, the cheeses that are on offer will be there [at the local supermarket]. Just as the leaflets are always a little bit compiled. There are the dairy products with yogurt, and there's the cheese." - P8

**Limitations of a Skill's Function Compared to the Device's Features** Particularly, participants were disappointed when a skill's functional scope did not match their expectations without further explanations. For example, some participants clearly expected to order REWE products via Alexa Echo Show and were surprised when they found out that this was not possible, especially because they knew they could order Amazon products. Similar encounters included managing shopping lists in Bring! and then being unable to purchase listed goods. In this case, participants tried to work around these limitations by trying at other skills which offered the desired functionality. In general, all participants expected Alexa to orchestrate all skills and support their activities alongside switching between skills, e.g. getting royalty points called 'payback' also at REWE and not only at real;-. Usually, the participants would activate several skills at once at home and expect them to complement and exchange information with each other:

> P7: "Alexa, please show the ingredients."

> Alexa: “Here are the: ‘ingredients’.”
> P7 *(confused)*: “These are not the ones I read to her.”

Yet, they did not immediately realize that this transfer was not implemented, because they first had the impression that ‘ingredients’ is the category under which all the ingredients collected in different skills are summarized. Instead, they accidentally activated Amazon’s own shopping list. However, missing indicators and feedback made it difficult for some participants to successfully track the behavior of Alexa and always recognize the switch to another skill:

> “I think I have picked the [recipe] here from Chefkoch. Should be in the computer memory.” - P8.

### 5.4.3 Visualize and Vocalize Information

**5.4.3.1 Interplay of Modalities and Their Perceived Added Value** We were interested in how the Alexa Echo Show can support domestic practices and individual routines and whether multimodal interaction can be used beneficially. The participants further evaluated Alexa Echo Show’s value by mapping their daily practices to the scope of Alexa’s skills regarding efficiency or convenience. Consequently, most participants compared the tasks with their domestic practices, as this helped them to understand the basic functionality and behavior of the system. Some participants did not perceive the additional display as beneficiary and wanted to use voice-first as primary interaction mode, or even questioned the rationale behind the additional display:

> “Well, but if I actually use Alexa, I would not want to wave around with my finger. I’d actually rely on Alexa.” - P9

Despite having the potential of a visual and aural in- and output channel, skills often failed to beneficially combine those. This resulted in interactions that were perceived as being too bothersome or lacking the visual feedback channel. Furthermore, Alexa encouraged participants to use voice interaction when it would have been easier to guide users to tend to the provided display and touch interaction. In the following, we use the example of creating and managing a shopping list to further detail our observations.

When operating a list, Alexa encouraged our participants to ‘scroll’ by voice. P20 was wondering what Alexa meant by suggesting scrolling through the list because they could not imagine how to do this with voice. Consequently, they scrolled the list via touch input. In contrast, P13 attempted to manage the shopping list via voice input and encountered respective issues:

> Alexa: “To select an entry, you first have to search for it by saying ’scroll up’ or ’down’.”

> P13: “Scroll down.” *(no reaction from the device)*
> P13: “Scroll down.” *(again, no reaction from the device)*
> P13: “Alexa, delete number 6 from my shopping list.”
> Alexa: “To select an entry, you first have to search for it by saying ’scroll up’ or ’down’.”

Most of our participants were unsatisfied with the lack of visual confirmation and representation of their items when accessing or managing the shopping list. Only a few participants claimed that they are not bothered by a voice-only shopping list, while others mentioned that they "think seeing is better than hearing" (P5).

**5.4.3.2 Quality of the Visual Content** In addition to the interaction modality, our participants were critical of the relevance of the information and content displayed on the device, since it failed to contribute to a more intuitive interaction. Consequently, most users considered the display to be quite useless at the current stage. One reason was that the visual representation of content violated some basic design principles for GUIs. Moreover, participants required a balance of listening, watching and reading. For example, P6 was frustrated because Alexa read out the content even though they had the screen to visually refer to displayed information:

> “I found it rather annoying because there was too much to read to me. So, I look at this device, that’s why I have this screen and then [Alexa] really doesn’t have to read every detail I see here.” - P6

When combining visual with aural interaction for providing information, our participants did not have a clear preference. as long as the output modality was aligned with the participant’s context and taking, for example, their distance to the display or the purpose or urgency of information into account. Contrary to voice-only devices, our participants expected a well-integrated display complemented with an IPA to use it as a standalone device without having to switch to a computer or their phones for getting additional information. However, Alexa Echo Show failed to fulfil this expectation due to the unused space and visualized too few choices at once on the screen. In particular, some information could be aggregated on the screen to minimize the effort to turn pages:

> “So far, I haven’t noticed that the [screen] has somehow been well-used, because especially when these steps are so insanely stupid, you can display three or four steps on a page and not just write a sentence.” - P6

The display was intuitively linked to touch control and seen as an opportunity to access more visual features. While knowing how to use touch control, the participants had issues with the software-based keyboard and the lack of visual cues to access or operate further features.

Therefore, they perceived the display as being rather useless. Some participants also doubted that the display was used to its full potential by several skills. For example, when browsing recipes, several skills, e.g. real;- or REWE simply presented mood images on the screen. However, our participants perceived this as a waste of space and wished for better integration of visual components in voice-based skills:

> "I have the feeling that some apps or skills have been squeezed into the Echo Show without being designed for it, simply because the screen is not used at all. (...) So, I don't need large images if I can't see the information I want to see." - P6

Although all users said that visuals are helpful, especially when it comes to shopping products, they also agreed that further improvements are needed. Unfortunately, keeping an overview of ingredients, products, and recipes was often inefficiently organized. The major point of critique, however, was the poor balance between auditive and visual information representation. P6 accordingly summarized their impression:

> "There was hardly any content. So this was clean in the (laughs) maximum sense. There was just nothing." - P6

## 5.5 Discussion and Implications

In the following, we summarize and discuss our results based on the observations described in Section 5.4. As reported by our participants, Alexa is perceived and treated as a standalone device, rather than acknowledging the orchestration between several 3$^{\text{rd}}$-party services. This mental model, however, results in several misconceptions about navigational hierarchy, the skills' scope of functionalities, as well as data model issues implemented by the individual skills. In the following, we propose design implications to support the future conceptualization and development of more user-friendly IPAs and their respective skills in the light of the current body of knowledge. We also identified challenges in IPAs and skill design that are already known, such as issues encountered by users regarding the robustness of voice commands [197, 221, 119]. Based on our lens of IPAs as an orchestrator of skills (see Fig. 2), we want to focus on new observations and what future designers of IPAs and skills could learn by that. We contribute with the following four key design implications: (1) contextual embedding of a voice-first design approach for skills, (2) differentiation between global (agent-wide) and local (skill-wide) contextual hierarchy, (3) scope of skills need to match users' expectations, practices, and preferences, and (4) the appropriateness of modality selection and agent behavior for specific activities.

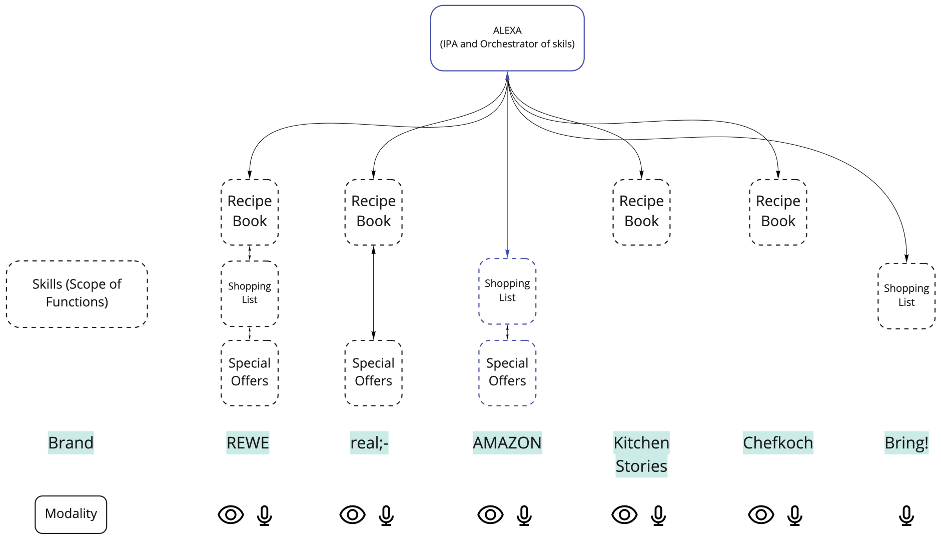


**Figure 2**
Alexa as an IPA orchestrating skills.

### 5.5.1 Towards a Voice-first Design Approach for Skills

Our results show recurring situations in which participants encounter issues caused by skills, which were presumably designed based on an existing concept for smartphone applications or websites. Yet, without altering data models and interaction concepts accordingly to the new device. Voice interaction requires fundamentally different application structures regarding presentation and data. We observed how our participants tried to reuse the interaction paradigms they already knew based on the 'original' application, e.g., the recipe website. Our observations of interacting with Chefkoch and Kitchen Stories further indicate that providers with atomic data and unified recipe structures can easily reuse them, whereas user-generated content needs to be explicitly cleaned and formatted for VUIs paired with a display. Further, conceptually thinking, websites and applications consist of several components that are interconnected to allow for smooth transitions between content. Therefore, we need to rethink content and functions from previous digital services as components that are made transferable to multimodal devices, if users desire to navigate by voice. Consequently, a voice-based IPA needs to take on the responsibility to guide users through the infrastructure and orchestrate between all components. Potentially, this could require more research and focus on conversational recommendations that are linked to the user account of respective services on further devices. Otherwise, we could also think of services as digital ecologies on different devices that allow only for an appropriate scope of functions and content. Although we did not evaluate IPAs provided by a smartphone OS, we see those issues and consequent implications adaptable to mobile phones, as designers of mobile apps have to make equivalent decisions on how far the content and data should be accessed and navigated via voice.

Following the concept of *mobile-first* vs *desktop-first*, we propose a *voice-first* design approach for the creation of skills: A skill creator's goal needs to be the design of compelling applications based on relevant user scenarios. At least, our results suggest that power users of distinct services or who have strong connections to brand or product experiences, will take longer to adapt to voice-first services. However, more investigations on their perception and understandings could lead to more design insights for an appropriate transition to voice-first. In this line, an application's complexity needs to fit both a user's routine and their task goal, supported by the beneficial combination of the available modalities [239].

### 5.5.2 Differentiation Between Global (Agent-Wide) and Local (Skill-Wide) Contextual Hierarchy

Our participants often lost their orientation, not being able to differentiate between the actions triggered by the skill or the IPA. They specifically had trouble keeping track of *Where am I currently in the application?* and *What does my voice command operate?* Similar to super apps [51], users expected the IPA to orchestrate its main functions and all activated skills. Due to the specifics of the speech, users perceived the interaction as one auditive stream of information [88] and oftentimes lost their awareness of the skills' scope. Whereas supper apps [51] manage to communicate the state and keep the conversation on track, Alexa Echo Show regularly confused the participants by not operating on consistent commands, for example, "undo" or "skip", as we detailed in Section 5.4.3. The use of “undo” is usually a crucial command to correct errors, but its inconsistency across skills leads to performance issues and abandonment of the skills [220]. Furthermore, our study revealed that the contextual hierarchy of global and local commands was not clear. Moreover, participants did not understand why interactional issues occurred because this distinction was neither explained nor communicated.

Therefore, standardization of voice commands across all skills might reduce errors and the unwanted quitting of skills, especially if the IPA better assisted users in not losing their touch to the skill they are currently operating. With previous research on IPAs reporting on interaction issues [53, 289, 60, 196], it is likely that IPAs that are based on similar eco-systems or business models, as Google Assist, might share related root causes of global and local hierarchy. While the Echo Show offering 3$^{rd}$-party applications, this might be still different for other eco-systems which albeit provide various skills or actions but have stricter use of language specifications for publishing, e.g. commands and activation words.

Over time, service platforms and ecosystems establish distinct conventions to allow users to navigate effortlessly, also when incorporating 3$^{rd}$-party applications. However, IPAs lack such infrastructural responsibilities and do not meet the users' expectations — one of the potential reasons for their abandonment [53]. Nevertheless, related work already reported that users refrain from using the full scope of functions offered by IPAs [289, 68, 197, 196], potentially due to a lack of personalization [53]. Therefore, allowing users to customize their

IPA, for example, through letting them choose their own activation words [53] or specific vocabulary for their daily routines might be perceived beneficial. While customization to define standard skills for certain activities exist, none of our experienced participants mentioned those settings to preconfigure the device according to their needs.

### 5.5.3 Scope of Skills Need to Match Users' Expectations, Practices, and Preferences

Skills expand the capabilities of the IPA and promise to further support users' practices. Although the tested skills matched the desired functionality of our participants, they perceived the implementation of the skills as rather disappointing: They expected the skills to be as thoroughly designed and implemented as other services they knew from the respective providers. In addition, and as described in Section 5.5.1, our participants expected the same range of functions implemented in a skill if a corresponding app or website existed. We think this was also caused by the lack of onboarding measures after a skill was activated for the first time; such a tutorial was only provided by real;-. Therefore, some participants were disappointed when data like specific recipes or functionalities like services to collect loyalty points were missing. Additionally, when participants reached the end of a skill's scope, they expected to switch effortlessly to other skills. We observed this, when participants expected to manage a shopping list's content via multiple skills or transfer a recipe's ingredients directly to a shopping list. Again, examples of supper apps [51] show how to make this possible and that it also represents a main benefit of integrated solutions. Although Grudin and Jacques [120] described IPAs as knowledgeable but shallow, we see attempts to account for the envisioned expansion of capabilities. Therefore, skills might not meet the user expectations of natural conversations yet, but those of utilitarian terms [61].

We, therefore, conclude that users benefit from basing a skill's scope on specific use cases, user routines, contexts, a clearly communicated skill scope and interoperability of all components. Finally, each skill should offer a tutorial to onboard novice users.

### 5.5.4 Appropriateness of Modalities and Agent Behavior for Certain Tasks and Activities

Currently, IPAs seem to only provide access to contextual information, but lack reliable interaction to effectively and efficiently guide users through information hierarchies. As a result, the discoverability of content and skills by our participants was limited. Based on our study, touch was predominantly used, to end frustrating conversations, also since users could not find a way to operate skills through directly interacting with the display. Drawing from the described observations, we argue that proper mode selection can only be done if the needs and characteristics of a specific task are both known and met. In line with the call for robust heuristics for CUIs [60, 220], we additionally suggest fitting skills to users' mental models of tasks, especially if they include a potential modality switch. In contrast to GUI, user control over speech-based systems is seen as rather limited due to its "invisible nature" [66] which makes it difficult to review and modify past actions [295], such as correcting wrong commands. Therefore, switching modalities is almost certainly required for some tasks. As we reported in section 5.4.3, our participants faced problems when verbally editing long lists of items because it required time to listen. Additionally, processing this amount of information increased their cognitive load to an extent where it might have negatively affected their performance [239] and lead to frustration. Even though our study design required to perform such actions which might not have been intended by the respective skills' creators, we see this example as an opportunity to demonstrate how users require to smoothly transition between modalities (see also Section 5.5.3). While today's IPAs do not yet offer activity recognition and can therefore only adapt to the context of use to a certain extent, it is even more important to allow users to actively switch input and output modes at any time. Further, switching modalities on mobile phones might need even more subtle consideration, as users expect those devices to fit all purposes at anytime. Subsequent research into the specific mobile applications might still uncover specific use that implies the use of voice, e.g. in the case of cooking or driving. Our case shows and exemplifies particularly the challenges of stationary multimodal devices.

Nevertheless, users must be informed if a skill requires a mode switch to function properly, in addition to keeping the application's context when a user decides to perform such a modality change. Unfortunately, we could observe that once our participants used touch to provide input on the Echo Show's display, they could not manage to pick up prior conversations and had to quit the skill. Building on that, modalities require respective signifiers to allow for intuitive interaction. However, the lack of both auditive and visual signifiers contributed to the users' loss of control and limited the discoverability of skills [223], and negatively affected our participants' orientation and their recognition of brands [351]. In our study, the burden of interpretation and establishing mutual understanding was shifted back to the users, since the IPA lacked proper feedback and a decent interplay and integration of visual and auditive signifiers. Here, especially, the design has a great potential to compensate for the lack of an IPA's intelligence in the interpretation performance [53] by building up dialogues and visual support and to increase the comprehensibility as well as to transport the emotions as for example in the inspiration.

### 5.5.5 Limitations and Future Work

In this study, we aimed to focus on qualitative information to understand why certain issues occur and how our participants explained their mental models of IPAs. Therefore, we refrained from conducting a dedicated usability study. Based on our user-centric view on how IPAs with a display can support daily routines in a household setting, we proposed several design implications to enhance the future interaction with such multimodal devices. While our implications target orchestrators of voice-based applications and 3$^{rd}$-party service and skill provider, we still lack knowledge about approaches to skill creation and design. Additionally, those implications might be slightly different for other IPAs, devices and ecosystems, as they might operate differently and offer a deviating skill-set. As we did not investigate different IPAs, we can only speculate at this point. Therefore, it is fair to state that our implications are preliminary and require both proper implementation and further evaluation, to prove that they are enhancing an IPA's user experience in general. Hence, we encourage researchers to conduct comparative studies with a larger user sample to derive more robust implications for these types of devices. Furthermore, our study focused on how German users interact with Amazon's Alexa Echo Show, leaving out perspectives from users of other ecosystems and cultural backgrounds. To address those limitations, further research should clearly focus on a culturally diverse user samples and consider varying contexts of use as well as language specific requirements. While our study focused on a specific routine - namely practices around cooking - we believe that our findings at least partially hold for other activities. We, hence, propose to focus on IPAs' potential to support a broader set of routines in daily practices.

## 5.6 Conclusion

In this paper, we presented a qualitative study investigating how users interact with Amazon's Alexa Echo Show, a standalone voice assistant paired with an interactive display. We were able to confirm that known interaction issues with IPAs, such as NLP robustness when it comes to recognizing and executing voice commands, is still a problem. Furthermore, our results show that the current integration of 3$^{rd}$-party services to increase the functional range of the IPA, so-called skills, does not meet our participants' expectations of flawless interaction while supporting their routines. Skills lack appropriate data models and do not make proper use of the additional screen to interact with or depict information. We reported that users frequently lost their orientation and had trouble associating voice commands to specific skills or global functions, as they perceived the IPA as a single device rather than a skill-orchestrating system. Finally, by applying our lens of super apps, we could discover further conceptual issues in the design of IPAs. Further we contribute to the body of existing research by proposing design implications enhancing the usability of IPAs, such as following a voice-first approach when designing skills, proper differentiation between agent-wide and skill-wide contextual hierarchy, required matching between skills' scopes and users' expectations, practices, and preferences, and appropriately selecting interaction modalities to beneficially support tasks and activities.

# 6 Morning Routines Between Calm and Engaging: Designing a Smart Mirror

## Abstract

Frequently the main purpose of domestic artifacts equipped with smart sensors is to hide technology, like previous examples of a Smart Mirror show. However, current Smart Homes often fail to provide meaningful IoT applications for all residents' needs. To design beyond efficiency and productivity, we propose to realize the potential of the traditional artifact for calm and engaging experiences. Therefore, we followed a design case study approach with 22 participants in total. After an initial focus group, we conducted a diary study to examine home routines and developed a conceptual design. The evaluation of our mid-fidelity prototype shows, that we need to study carefully the practices of the residents to leverage the physical material of the artifact to fit the routines. Our Smart Mirror, enhanced by digital qualities, supports meaningful activities and makes the bathroom more appealing. Thereby, we discuss domestic technology design beyond automation.

## 6.1 Introduction

In recent years smart home systems to save energy, increase security, and enable (self-)monitoring were researched and developed [126, 148]. For most parts, the current Internet of Things (IoT) is built to collect data and automate routines [322, 47, 356]. By making the gathered information accessible to households, typical IoT consumer technology design shall facilitate behavior change or real-time reactions to unusual events. Additionally, Intelligent Personal Assistants (IPA) are increasingly integrated into speakers or ambient displays to allow for 'natural' interaction with all IoT appliances [6]. That falls in line with Weiser's vision of calm technology [346] with technology 'disappearing' and little to no digital interruption.

However, such design credo ignores user expectations of engaging and exciting interactions, for example, when talking with IPAs [53, 60]. Mostly, building close relationships with technology fails as users desire true conversational interactions going beyond short and single commands [53]. Besides IPA control interfaces, other work even indicates that people fear becoming passive and lazy in fully automated home settings [210, 5]. This lack of practice engagement and missing meaningfulness throughout IoT interaction leads to limited long-term use of home IPAs and even non-use [53, 197]. Still, the concept work of smart home artifacts is often technology-driven with the main purpose to conceal ambient displays by neglecting the variety of domestic needs and values [7, 8, 54]. Instead, meaningful qualities that traditional artifacts inherit, should be further digitally extended [123].

To explore the potential of making traditional artifacts interactive, we investigate the design space of a Smart Mirror. Thereby, we followed a design case study approach as proposed by [362]. First, we studied entangled morning and evening routines in and outside the bathroom by a focus group of seven and a diary study of ten participants. Based on the material, we developed a modular concept as a mid-fidelity prototype. Lastly, we evaluated the mirror design with five participants in their bathrooms.

Our findings indicate that the design trend for "optimization" of domestic routines limits the perspective on valuable smart artifacts. Our prototype offers an alternative design for pleasant interactions that fit personal and steady as well as rapidly changing routines and needs. However, many of them create space for self-care or conscious moments of reflection or creativity. Engaging applications like embedded in our Smart Mirror may support those pleasant activities.

## 6.2 Related Work

### 6.2.1 Smart Artifacts Between Automation and Control

From a traditional perspective, smart home technology has been mostly associated with optimized, efficient routines and autonomous decision-making of the system [322, 48, 126]. This includes installations of automation infrastructure to save energy, to increase safety, or to enable (self-)monitoring [47, 48, 148, 63, 339]. According to [210] the perceived benefits are "small conveniences rather than substantial support for routines". Previous work [210, 5] shows that people fear such technologies to deprive them of the activities they enjoy and, hence, make them passive and lazy. Furthermore, frequent notifications contribute to a constant distraction and reduce well-being [359].

Smart Speakers, Displays, and Mirrors are frequently introduced as smart home control interfaces [60, 8, 7]. Following the predominant design paradigm, their main purpose remains to control lights and music, inform about weather conditions, or set reminders [6]. However, [53] shows that users expect those devices to interact intelligently. This mismatch disillusions long-term users, subsequently adapting their language and expectations [53, 6]. One reason devices are not becoming substantial is the lack of engaging interaction and greater support for daily routines. Pleasure is limited to colorful mood-setting, light controls, or connected entertainment devices [310, 149].

Similarly, most studies investigating the design and use of smart mirrors focus on ambient information access [7, 8, 54, 339, 100]. They often lack the enhancement and extension of their physical properties such as the mirror surface, but merely serve to mask built-in technology. Persuasive mirrors [225] tend to overemphasize behavioral change for long-term interactions and objectives, while meaningful applications can also arise from sporadic interactions.

### 6.2.2 Towards Engaging Artifacts

By recognizing the current downsides of the predominant design paradigm, various researchers proposed directions for a future beyond automation and control [86, 269, 75, 310]. There is a chance to understand the home as a design space inspired and shaped by various interactions and activities between residents and artifacts [171, 57, 240]. Here, [105] argues, "unless we start to respect the full range of values that make us human, the technologies we build are likely to be dull and uninteresting at best, and dehumanizing at worst". In particular, when we treat domestic practices with the same optimization approaches as the workplace [69, 126]. Similarly, [75] propose placing a stronger effort into conceptualizing and exploring the look and feel of alternative visions of co-living with smart IoT. Therefore, design approaches should leverage the range of activities performed in the home rather than decrease their relevance through automation. In this light, [132] argues to focus on positive activities. Therefore, the level of interactivity does not necessarily have to be reduced in favor of efficiency [86, 75, 132], but enhanced towards more enjoyable interactions.

Verbeek's [329] notion of things that 'act' allows following this perspective by recognizing the values and inherent attributes of the artifacts as actors. These properties "enable and constrain certain ways of interaction simultaneously."[98] and thus, allow the building of close relationships between objects and residents through greater engagement and personal interactions [151]. A mirror surface, for example, is appropriate to display content but simultaneously confronts people with self-reflection as they observe themselves [215]. Onward, it may also support workouts [125] or even art [144].

Hence, we need to explore how to create interactive resources for engaging experiences that support currently performed activities [269] by understanding the context and already established material of the domestic practices. We thus aim at better understanding what it means to shift between calm and engaging experience and how to design for more well-being in Smart Homes.

## 6.3 Design Approach

Following a user-centered design approach, we conducted a Design Case Study by [362] to align the design of an interactive mirror with the needs of potential users. At first, we conducted a focus group to discuss the actual use, meaning, and entangled practices around the mirror to determine potentially engaging design opportunities. Due to the primary use of mirrors in the morning and evening, we continued with a diary study of according routines and follow-up interviews. The results of our formative study led to a conceptual design of four separate digital applications later embedded in the artifact. Finally, we evaluated the prototype in a Wizard-of-Oz study [366].

### 6.3.1 Focus Group

The focus group aimed to explore the meaning, actual use, and activities surrounding mirrors in everyday life. Therefore, four female and three male participants, aged between 26 and 29 years, were recruited by snowball sampling. We decided to foster discussion by inviting three early adopters who can weigh in their experience and curiosity towards consumer electronics and four technology critical and hesitant adopters. The discussion was led by the host asking guiding questions but otherwise remaining silent. After a brief personal introduction, participants shared their estimated time per day in front of the mirror and situations when and where actively using the mirror. Thereby, the most commonly reported practices involved the bathroom. Afterward, mirrors as home materials and goods were discussed. We intended to encourage reflection of personal experience and interaction with the traditional mirror to explore new design possibilities for computational properties. Finally, each participant sketched on paper their personal vision of an ideal mirror with potential applications and desired interaction. The discussion was audio-recorded, transcribed, and thematically coded by two researchers. Afterward, we discussed the themes within our research group, likewise all participants' drawings [32].

### 6.3.2 Diary Study

To gain a thorough understanding of the qualities of a traditional mirror and associated routines, ten participants shared information about their everyday life for seven days within our diary study [28]. The sample was heterogeneous considering prior technical knowledge, marital status, occupation, living situation, and experience with digital assistants. It was recruited by snowball sampling and aged between 21 and 33 years. Six participants had a significant other, with three of them living in the same household. The others lived alone or shared an apartment. The type of education or occupation partially structured their everyday life: Six employees, with one working frequently from home, one student, one pupil, one freelancer, and one mother occasionally working as a freelancer.

We used EthOS[1] to record everyday events that participants logged via their mobile phones, and supervisors were allowed to view, sort, and code entries simultaneously. The first question required a photo response, and the other questions alternated between descriptive free-choice media and forced/multiple-choice options. A total of five questions had to be answered descriptively in the morning and evening, and three multiple-choice items in the morning and two in the evening. Besides automatic notification of any changes, the supervisor sent emails twice a day as reminders to the participant.

Afterward, in-depth interviews (70 minutes on average) addressed possible ambiguities and specific questions on occurring events. As the recordings were limited to a one-week diary study and represented just a fraction of daily life, we aimed to reflect with the participants on

[1] www.ethosapp.com

their perception of their behavior, such as the general handling of digital devices, the corresponding applications, and its meaning to them. The mirror was discussed as an interactive artifact in domestic spaces between relaxation and activity within evening and morning routines. Thereby, reported photos and further media enabled us to ask more detailed questions about context-related activities such as daily planning, bathroom activities, relaxing routines, morning motivation and priorities. Finally, the participants were asked to express their ideas and criticism on the applications, interaction, and the use of a smart mirror. Particularly, we aimed to discover differences between people as well as deviations in the same person [28]. Therefore, daily reported enumerations, descriptions, experiences, and final interviews were coded and analyzed for similarities, differences, ambiguity, and needs.

## 6.4 Contextual Analysis

### 6.4.1 Expectations of a (Smart) Mirror

The participants estimated their mirror use between five and 30 minutes, on average 18 minutes a day. All participants owned a bathroom mirror which they referred to as the principal mirror of use. The actions performed in front of the mirror range from the last glance before leaving the door to engaging interactions like conscious personal care. The bathroom itself represents for most of the participants the most private and intimate space in the home. Within this context, the mentioned media applications are strongly entangled with personal morning and evening routines, as the most frequent hours spent at home. Most of the participants desire functions related to infotainment and organizational tools. The analysis of the drawings implies that all of them request effortless syncing of their favorite smartphone applications with the mirror, besides monitoring home appliances. Some participants see great benefit to watch make-up tutorials on a Smart Mirror. Presenting the drawings of their 'Dream Mirror' revealed new ideas and thus mutually influenced the desires and inspirations of all. Hence, alternative scenarios encouraged the evolution of further needs that had not previously been thought of by all [269]. Although many of the described functionalities would require cameras and microphones for implementation, every participant had privacy concerns regarding smart home systems.

### 6.4.2 Morning and Evening Routines

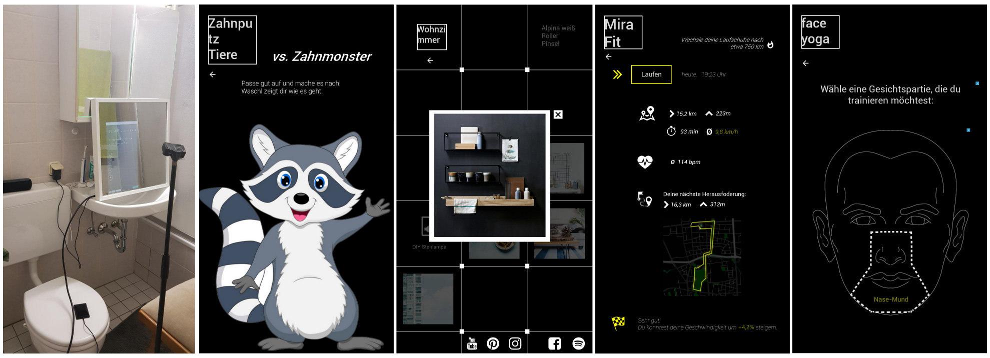
The documentation of daily digital activities, the interaction with physical objects, and the associated significance for the participants provide information about the interrelated factors that influence well-being at the corresponding time of the day. Both individual moments and long-term use can provide context-based personal goals and values.

Spending time in the bathroom ranged daytime-specific from five to 15 minutes for a short stay and 20 to 45 minutes for more time-consuming practices. The average time for each participant per day turned out to be quite similar. All participants expressed that their bathroom design has a considerable effect on their well-being and thus on their stay. Therefore, they had hung up personal pictures or photos and set up decorative elements such as plants. Lighting design, music system, bathtub, and photos contributed significantly to a pleasant bathroom atmosphere. The weekend resulted in several short visits to the bathroom as there was no time pressure compared to workdays. The sequence of activities differed between participants, but brushing teeth or drinking coffee in the morning were usually among the first after getting up. Longer stays usually involved showering with body and face care. Dental care usually was done twice a day and took a planned minimum of two minutes. Meantime, the activities performed consisted of looking in the mirror, doing nothing, seeking engagement or entertainment in or outside the bathroom. Some noticed their tired face or checked it for health in general. In the evening, all female participants followed their facial skincare routine.

Participants considered a morning atypical as soon as something unforeseen had to be done, thus increasing the time pressure. The same applies to difficulties getting out of bed or being sick. One participant structured his morning with a mobile app that was designed to encourage good habits and included a checklist to do so. Another participant started to wake up by interacting with his mobile phone, while another one often laid down for a few more minutes to think about the day ahead. Many of the participants depended on public transport and therefore always kept track of time. Daily planning was sometimes omitted by those who had structured days and hardly required any additional preparation for work. Otherwise, participants intended to do the planning of work tasks in the office before leaving. Depending on the evening before, ‘morning activities’ could last the whole day or until leaving home. The use of reminders involved only cases of unusual events or for irregular notes like bringing musical instruments or sports equipment to work for after-work events. The morning routine at the weekend could no longer be recognized as such, as most of the participants started the day without any time pressure. Shortly before falling asleep, many of the participants reached for various media such as videos, books, music, or devices to browse the Internet. Media activities can generally be categorized as follows: Social media, news and communication, entertainment, online learning and tutorials, health and well-being, shopping and renting, dating, and smart home appliances. For more complex tasks, participants preferred a bigger screen size and the appropriate interaction style.

## 6.5 Conceptual Design

Previous results show that even if established routines exist, participants carve out time for (self-)reflection, conscious personal care, or enjoy moments of doing nothing in specific. Daily planning was either done during the working time or only in case of unusual events. In contrast to previous research, we wanted to go beyond the design for optimization and efficiency and focus on conscious and active moments of interaction and experience. Thereby, aesthetics, personalized design, and well-being strongly influenced the stay in the bathroom. We consolidated the gathered data, needs and actions, and used a scenario-based design ap-



Passe gut auf und mache es nach! Waschl zeigt dir wie es geht.

Wechsle deine Laufschuhe nach etwa 750 km

Deine nächste Herausfoderung:

Sehr gut! Du konntest deine Geschwindigkeit um +4,2% steigern.

Wähle eine Gesichtspartie, die du trainieren möchtest:

**Figure 3**
Evaluation set-up and exemplary prototype screens: Tooth Brush Animal, Instant Idea, MiraFit and FaceYoga.

proach to develop our mid-fidelity prototype. As follows, we will outline our general concept and the prototype in detail.

The concept entails four main applications and a home screen with basic functionalities for communication, connected devices, and reminders (3). A personal agent enables navigation through several applications and functions to facilitate seamless interaction in the whole room. Depending on the application, the agent changes roles supporting the media style, e.g., the character of an animated animal. Our heterogeneous samples showed clearly that future designs have to be customizable. Meanwhile, our concept focuses on engaging support for the activities at hand, situated well-being, and (self-)reflection. The four proposed applications build a modular foundation for various engaging features and interactive media elements adapted to the needs of the residents and the specifications of a mirror.

## 6.6 Mid-fidelity Prototype

As follows, we built a mid-fidelity prototype to illustrate the possible look and feel of a Smart Mirror in reality (3). Therefore, we examined the application ideas to determine their feasibility with respect to the technical limitations of the hardware elements and the test environment for evaluation. The central part of the smart mirror is a light monitor display concealed by a spying mirror glass as an interaction surface. This surface is partly translucent, hence, hiding everything dark. We tried to provide an experience as realistic as possible regarding the interactive digital elements and current hardware options. The interaction comprises voice in- and output as well as visual support. Icons and text prompts provide an overview of the application and support menu navigation. The size and distance of the visual elements are proportional to the size of the display, taking up about half of the mirror surface. The arrangement of the elements attempts not to interfere with the person's mirror reflection. By triggering an application, the content extends across the entire display as needed while enough mirror surface remains usable as such.

**Home Screen.** Organization and communication tools have clear value for domestic life.

Hence, for completeness, we decided to incorporate visual signifiers that illustrate further tools. As with conventional end devices, the screen functions as an overview of the installed applications that are represented by icons. Those are accessible by saying the voice command 'Mira, please start application xyz'.

**Face Yoga.** This interactive video tutorial is based on yoga practices to train and relax facial areas and support mindful recreation. Short audio and video instructions provide exercises for different face parts. They offer personal care for personalized time budgets of two to seven minutes, creating opportunities to incorporate more active time with and for oneself. Calm music and wording contribute to a reflective atmosphere by building on the context of Yoga. The agent acts as a guide and describes further steps as well as the flow of the exercise. The user can check the correctness of movements simultaneously in the mirror. For demonstration purposes, we have currently provided offers such as anti-aging and relaxation exercises only. Finally, these should be easy to integrate into personal care routines and create moments of conscious interaction and reflection.

**Tooth Brush Animals.** In this application, the agent acts as one of three animated animals to motivate and teach children to brush their teeth properly. Therefore, the agent transforms from a calm background assistant to an active coach and has distinct characteristics to build a trustful and engaging relationship with the child. Further, the animal tells a story about 'little tooth monsters' that try to attack the teeth and can be defeated only with the help of proper tooth brushing techniques. An animated dental model shows the correct brush movements in the oral cavity. The embedded gamification approach aims at additional motivation by offering further animals to unlock. This approach can be applied and modified for different topics relevant to children and to support parents in child care. However, this is a way to create more well-being in bathrooms which often seem sterile and not fun espacially to children.

**Instant Idea.** Many of the best ideas and creative moments arise during a moment of relaxation and non-activity of the brain [170]. Additionally, people take active time to think of the day ahead. Instant Idea shall support users to capture and pursue spontaneous ideas right in the bathroom, e.g., by doing a voice memo or an image search. Moreover, every media format and platform, e.g., videos, screenshots, or tweets, can be saved for a future purpose and processed on other devices. It is possible to insert and compile the content into a personal grid of inspirational ideas and quotes, allowing users a versatile combination of the collected material to develop new ideas and concepts. In general, this application supports thought activities and creative moments by enabling users to capture and structure their thoughts.

**MiraFit.** MiraFit imports and visualizes data like current activity results and goals collected by the users' preferred mobile fitness application or tracker. The aim is to support personal care and self-reflection on physical goals as well as sports habits. The mirror proactively visualizes information and acts as a coach with further data-based advice. For example, motivational quotes from successful athletes will appear or new challenges are proposed. This application is timed to suit the users' post-exercise needs while they are following their care routines. In particular, the main results should be visible at first glance and engaging to users for future habits, exercises, and well-being.

## 6.7 Evaluation

We conducted a heuristic evaluation to explore the potential value of the concept and discuss future design implications for smart artifacts. At least two of the four applications matched the participants' needs and goals for personal care and well-being in the bathroom. We recruited our participants by snowball sampling with an average age of 28.4 years, ranging between 23 and 32. Two of them lived alone in a one and two-room flat, and one shared a two-room flat with a co-inhabitant. The other two participants are married and live together with two little children, three and six years old, in a three-room flat. All participants had some experience using mobile voice assistants.

For an authentic atmosphere and personal experience, we conducted the evaluation in the bathroom of the participants. The mirror prototype 'Mira' was placed on the washbasin in front of the actual mirror (3). The test leader and assistant observed the interaction next door on a live video stream. The test leader acted as 'voice assistant' within the Wizard-of-Oz [366] scenario set-up. Therefore, we equipped the bathroom with a microphone and a small speaker. The participants used their voices to command actions. The test leader executed them by clicking on the according elements in the digital prototype transmitted from the laptop to the Smart Mirror display. Each participant received two scenarios and a corresponding task. Meanwhile, the actions were recorded on video, and observations were noted. Afterward, participants had additional time to explore the rest of the content freely. The participants were asked to think aloud during the whole session but needed to say 'Mira' as an activation word to use voice commands. Otherwise, the test leader would not respond. General questions that emerged during the test were not answered until the end unless the participants had specifically asked Mira for such information. In total, one session lasted between 45 and 60 minutes. The videos were transcribed, coded, and deductively classified in MAXQDA [32].

### 6.7.1 Findings

All participants had a positive first impression, like P1 noted: "I actually thought it would look more 'Do It Yourself' and not so professional". They described the handling as intuitive and well-structured. All of them agreed that the applications added value, but the better they corresponded to personal practice, the greater the enthusiasm: "Fitness or relaxation helps me to relax in everyday life, it could increase my quality of life" (P3). As P5 elaborated: "The idea with the toothbrush animals is great, Instant Idea is very good, FaceYoga I don't know". Most of the participants successfully completed every task and handled the fictive voice control well. In particular, three participants completely blanked out the test leader, fully engaged in the interaction, and explored all applications via voice command. Both during the test and in follow-up interviews, the adjectives mentioned, such as ‘presentable’, ‘innovative’, ‘likable’, ‘pleasant’, and ‘motivating’ indicate a positive experience.

**6.7.1.1 Home screen** Participants particularly embraced the displayed clock and reminders on the home screen, e.g., taking their sports equipment with them, since they had no time indicators or note board in their bathrooms. Offering clothing suggestions in the case of rainy weather was perceived twofold: “I will make my own decisions for myself, just take a jacket or umbrella is fine” (P5). He was deliberately refusing to be patronized by technology. Further, some participants would have liked more visual indicators for possible voice interaction because of the unfamiliarity with it.

**6.7.1.2 FaceYoga** The opinions on this application differed. Regarding design and interaction, the participants described it as entirely positive, supportive, and pleasant: “It was clear which parts of the face could be selected. I also found it good that the voice guided me. I also liked the fact that it asked me if I wanted to do another exercise.” (P3). In terms of content, however, only two of the participants would repeat the exercise and one would try out different exercises first because this particular session was not effective. Several participants emphasized the benefits of demonstration, an easy way to simultaneously join the actions and correct oneself by using the mirror: “It’s super intuitive because you can’t go wrong with it.” (P2). In the beginning, P4 was irritated by the simultaneous dubbing and texting of the instructions because of a mismatched timing: “I would not have needed the written instructions because Mira explained it to me”(P4). P1 would prefer effortless switching between speech and text, as he sees an advantage in both. Besides, he experienced difficulty in his hand-eye coordination between watching the video and checking his movements in the mirror all at once. Therefore, he suggested that the video should overlap with his face in the mirror. Overall, while there is a value for personal well-being, more individualization is desired. Text, sound, and image should be more balanced to ensure smooth interaction.

**6.7.1.3 Tooth Brush Animals** All participants watched the animation, and two of them got engaged. The general impression was positive, and everyone could imagine children enjoying the application. P5, a father himself, indicated the mirror “anchors learning where learning takes place”. Gamification purposes like unlocking further animals were well received and created immediate engagement: “Mira, which animal suits me”(P1) or “Can I create a new character there?” (P5). The video is well suited to develop a sense of time, and the raccoon is, in any case, a positive factor to increase motivation. However, most of the participants criticized the fast movements in the cartoon. P5 mentioned limits to check his teeth simultaneously in the mirror at this speed. Besides, he added that the final check of the child’s teeth and responsibility still lies with the parents. Thereby, it would not save time. However, even if he did not buy the mirror just because of one application, he would install it if he already had one. Yet, some of the participants had parental concerns about exposing children to further screens. One parent (P5) was excited “to see how they do it with the mirror and imitate. How much they stick to it”. The other parent (P4) added that “If they are more grown-up and are allowed to brush their teeth themselves at lunchtime and without supervision, that would be something”.

**6.7.1.4 Instant Idea** Despite initial insecurities about the concept and application, all participants agreed that the concept was valuable and fit their routines: “I really believe that I would use it because I always use notes to write something down. Just like that, if I think of something, I would write it down briefly. And you can hold on to it without searching for my mobile phone with my wet hands, and I can do other things on the side”(P5). P4 recognized its practicality, e.g., while brushing teeth or doing make-up. The visualizations of the sequences were very authentic in their functionality, whereby the prototyped interaction caused confusion. However, one reason was the simulation, where certain options had to be prepared in advance and some restricted in use. Therefore, one participant suggested having animated hints, for instance, to record his voice. Besides, he wished for references to the sources of the filed media in the future. The social media links were only noticed on a second look but tended to be positive. One of the participants wanted to sort his stored content by link categories like ‘living room’ (P4). Despite the initial difficulties in interaction, this concept has the potential to support reflective moments and creative thinking.

**6.7.1.5 MiraFit** All participants liked the well-structured fitness results and suggestions by Mira. They emphasized the automated synchronization of data and the mirror application to reduce additional effort. Equally, after exploring the training advice on the mirror, they would like Mira to send the information to their phones or fitness tracker. Besides, one participant asked for an automatic calendar entry for the next run. P1 noted that he is not familiar with the displayed times on the mirror, and he is expecting classification and interpretation by the agent. All participants emphasized the importance of context and timely suggestions, for instance, depending on different times for workouts, as P2 remarked. Besides, recommendations to buy new running shoes after a specific number of miles were well received, with P4 expecting to get this information timely and simultaneously some links for direct purchasing. All participants asked for valuable advice and information and emphasized the importance of timing to engage with the data and the agent.

**6.7.1.6 Impact on the Atmosphere and Well-being** Four participants described their bathroom as a very intimate retreat, where they particularly want to feel calm and cozy: “It’s a very private room, you’re usually alone there” (P2). Further, they repeatedly emphasized the positive impact of the mirror on the atmosphere in the bathroom: “I find it very user-friendly and a bit like a girlfriend in the bathroom (...) With the mirror, the bathroom would no longer be so sterile and cold, but cozier.” (P3). P4 added the positive effects of speech:

"So that makes it more human, of course, because of the voice." However, P5 encountered: "I don't know if you need a name for the mirror and if it has to speak, perhaps it would be enough if only I would speak and then I get the feedback on the screen" (P5). He did enjoy the interaction, but sometimes visual feedback would be sufficient to engage with the content and activities. Without considering additional effort to clean fingerprints off the surface, he also would like to shorten some interaction paths by touch. In contrast, P3 explained that speech particularly fosters engagement and motivation.

Although the participants enjoyed engaging with the mirror, they all had privacy concerns. Even without an integrated camera in this prototype, they mentioned, it would make them feel uncomfortable. Likewise, they were concerned that the sounds of the toilet might be recorded and distributed. Therefore, some participants suggested mechanical features like a flap to blind the camera and preferred a self-determined control. The same goes for switching the microphone on and off, as a digital marker, e.g., light still leaves them suspicious. Besides those reservations, the participants valued most of the applications and made design suggestions as taking selfies in the bathroom without considering prior stated privacy concerns.

Some participants speculated on more design ideas to enhance their well-being. Aesthetics contributes equally to a sense of well-being as the applications themselves. P2, for instance, imagined an effective weather display by letting rain run over the mirror or a sunrise. Similarly, P1 suggested the mirror simulates a real window or a mood light to feel more comfortable in his small and window-less bathroom. That may also lead to spending more time in this particular room, in general. All participants already listened to music frequently, and some mentioned watching music videos as an additional benefit. Although spending most of their time alone in the bathroom or helping their children, P5 emphasized that this artifact might also impress and entertain friends and acquaintances at their visit: "It is a luxury item that is not only beautiful but also has a benefit. (...) It is also a wow factor for guests." (P5). However, P2 wished for a 'calmer' design, which reminds her less of technology like the mobile phone. The clock was a little too big, and she associated the functions of reading emails or getting messages on the mirror with her working day ahead. She would prefer to hide these functions and displaying the watch in an 'analog' design on the mirror.

**6.7.1.7 Fitting the routines** The results show that participants expect a personal fit to their habits and time-critical events. For the latter, one participant (P5) particularly described a stressful situation storming into the bathroom and handing over several tasks to the mirror. In this scenario, the agent has to react quickly and send, for instance, a voice notification to a friend for his 15-minute late arrival. Besides, participants reflected on their daily routines and possible fit of the applications: "In the morning, the applications that I tested, like FaceYoga. And in the evening perhaps rather as a little toy and for entertainment. And something like the news I would watch at noon. However, actively I would use the mirror in the morning and evening" (P4). Likewise, P2 added that this mirror might support a relaxed and organized start to the day: "You feel more organized, you do things that you would do anyway, and you get information. I would feel more comfortable with it." (P2). She also stressed that she would use beauty advice for skincare and make-up and preferred motivational content for the day. For building healthy habits, P3 emphasized the benefits of embedding the mirror in the bathroom and the immediate use: "I always miss to do the relaxation exercises, but if it's right in front of the mirror and you're right there, then you do it." Usually, the first thing she does after coming home from work is going to the bathroom, so she imagines starting an application immediately while drying her hands.

## 6.8 Discussion

We want to discuss the main findings of our design process in light of the design space for engaging interactions [268] and leveraging properties of traditional domestic artifacts [329].

### 6.8.1 Traditional Artifacts Extended

So far, research treated mirrors and ambient displays as very multi-purpose, public furnishings and artifacts for all household residents. Hence, they have been usually assigned the task of communication and coordination work. Ambient displays have traditionally been developed for public spaces to disseminate information widely and make it accessible to all. For the most part, enhancement or intelligence of IoT artifacts has been understood as the need to hide technology or visible aspects of domestic technology in everyday objects. As a result, the original meaning of the object and its inherent qualities, such as the mirror surface, and the primary moments of situated interaction are insufficiently considered. Moreover, when interacting with technology, the technology's need to communicate organizational information, for example, takes a salient role, forcing the residents to immediate reactions rather than supporting their environmental needs associated with the mirror and space. The spatial design of the bathroom impacts personal well-being substantially. [310] show that aesthetic and ambient features in the home are as important as the technology itself and lead to more pleasure. Therefore, the object and its properties carry well-being, either as a traditional material or digitally enhanced by applications. The qualities of the artifact enable and constrain the inherent interaction and expressiveness [123, 98]. Traditional mirror reflections shift the focus to more self-reflection, and with digital qualities, it is now possible to engage in active and reflective ways. Thereby, a digitally enhanced mirror might actively offer space and time for calmness and more engaging experiences in the "currently doings" [268]. The same surface might constrain the usefulness of some applications like the calendar in the bathroom and simultaneously be a valuable feature on a decorative mirror in the living room. With an iterative design approach, we were able to uncover actual use and entangled practices of the traditional mirror at home and show how to center those in the further design development considering the constraints and opportunities of the material. The examination of the social practice in which the material encounters meaning and the potential for use helps to re-contextualize the purpose of digitization and visualize the vital qualities of the artifact.

Our approach is not limited to mirrors but emphasizes exploring artifacts in their original embedded use to integrate technology purposefully and open up new design perspectives. Therefore, the main quality of everyday artifacts should go beyond concealing technology and find the natural fit by leveraging inherent properties and affordances. There is a potential to carefully extend properties digitally that build on prior structure, use, and desires and see IoT as active and embedded contributors to more well-being in the home.

### 6.8.2 Adaptive Resources for Action

We withdrew to condense the needs of our participants to an average user to avoid the 'One-size-fits-all' design paradigm. Leveraging the design space of the bathroom and traditional mirror, we present four applications that promote and inspire mindfulness and well-being in the home, aligned with the call of [75] for alternative IoT concepts. We based our concept on the engaging interaction between inhabitants and their artifacts, offering resources for action to find substantial and joyful support for their routines [211].

Concise moments and activities define the potential value and support of the technology for everyday life [300]. Our diary study shows the frequent media use in the mornings and evenings. Yet, we can observe participants attempt to integrate time for (sub)conscious reflection, self-care, and to establish enjoyable or healthy habits in general. In contrast, prior studies often neglect the variety of needs that can be projected on one artifact or the entanglement of different practices associated with one room or artifact. Those systems tried to enhance well-being by more automation of tedious tasks or processes like regulating heating [148] that not primary focus to promote joyful interaction but instead passive and peripheral information consumption [6]. With our empirical studies, we could reveal the entanglement of media use with the variety of morning and evening practices, pointing to different phases of calmness and engagement that personalized technology has to consider. This also extends to the investigation of the personal relationship between inhabitants and their objects in use. Regular encounters that involve memories, engagement, and experience create personal value and strengthen the relationship with the object, leading to appreciation and acceptance of the technology-enhanced artifact as well. However, our prototype shall enable humans without strongly intervening or patronizing, yet offer resources for engagement [316]. Users value a variety of unique applications to choose for their individual purpose and might build close relationships with the agent if their needs are taken seriously by design [240]. This will need long-term investigation of said relationships to understand how more IoT can, for example, live up to the expectation of being personal.

A thorough investigation of domestic practices with a central view on the material and respecting the former object relationship contributes to the creation of personal value within the adoption of the interactive artifact as a whole. Therefore, we need to find a balance between automation and engagement by offering adaptive resources to a variety of needs and connecting existing activities and objects.

### 6.8.3 Rethinking Productivity

At the beginning of the broad implementation of technology in homes, practices were investigated by means introduced to study workplaces, and success was determined by increased values of efficiency and productivity [69, 126]. Yet, we have to rethink the value of efficiency and productivity in smart domestic environments [76, 69] and their meaning to the inhabitants.

Time-economic advantages exist and can reduce stress by proposing an efficient structure or overtaking tedious, previously manual tasks. Yet, inhabitants might not experience this as a value because they do not mind, e.g., opening windows by themselves or they want to make own decisions. Consequently, they still might not perceive a technology dictating the daily structure and which is concealed by daily objects as calm. Calmness emerges from the absence of distraction and fitting interactions between inhabitants and artifacts. Ambient access to information does not increase efficiency necessarily when further activities like self-care or creativity are interrupted. For example, information retrieval in the morning might even produce stress by displaying work messages. Therefore, an alternative approach might be the active support of moments that often remain invisible to technology and unconscious to inhabitants.

Additionally, users fear becoming passive and lazy in the opposite of being productive, when too much automation is implemented in their homes. Understanding that being active equals not always being productive, we can move towards the design of artifacts and interfaces that promote engagement which is welcomed and desired. Productivity is often linked to specific goals and tangible results, whereas being active can also be associated with mindful experiences in the moment, e.g. self-reflection or self-care. Moreover, being productive can be understood as being active and engaged in favorite activities. Accordingly, technology should instead foster the reallocation of resources like time and space to more meaningful engagements. Tools for more self-reflection and mindfulness help to increase the productivity of the inhabitants throughout the day. Finally, our work enables users to implement more positive activities in their daily routines and establish desired self-care habits.

Finally, the properties of artifacts are appropriate to resolve the contradiction of calm and engaging by rethinking the values of efficiency and productivity. Therefore, we need to design beyond the automation of routines and control of smart appliances [75, 310] and consider which spaces in the home are appropriate for coordination and communication work and which are used for calm and mindful interactions.

## 6.9 Conclusion

Inspired by the idea of IoT artifacts going beyond efficiency by digitally extending the qualities they already inherit, this paper presents a design case study for a Smart Mirror that supports activities and is easy to integrate in everyday life. Our findings indicate that the design trend for ’optimization’ of domestic routines limits the perspective on valuable smart artifacts. Moreover, our ’Mira’ prototype offers an alternative design for pleasant interactions that fit personal and steady as well as rapidly changing routines.

Our research is limited by the small number of participants in the evaluation and selection of the sample, which should be broadened in future work. Moreover, we can only speculate about the design of other artifacts because they are determined by their inherent properties, still our results clearly show the need to investigate a variety of IoT artifacts. Further research should focus on digital enhancement of traditional artifacts and purposes for well-being beyond automation.

# 7 Trust Your Guts: Fostering Embodied Knowledge and Sustainable Practices through Voice Interaction

## Abstract

Despite various attempts to prevent food waste and motivate conscious food handling, household members find it difficult to correctly assess the edibility of food. With the rise of ambient voice assistants, we did a design case study to support households' in-situ decision-making process in collaboration with our voice agent prototype, Fischer Fritz. Therefore, we conducted 15 contextual inquiries to understand food practices at home. Furthermore, we interviewed six fish experts to inform the design of our voice agent on how to guide consumers and teach food literacy. Finally, we created a prototype and discussed with 15 consumers its impact and capability to convey embodied knowledge to the human that is engaged as sensor. Our design research goes beyond current Human Food Interaction automation approaches by emphasizing the human-food relationship in technology design and demonstrating future complementary human-agent collaboration with the aim to increase humans' competence to sense, think and act.

## 7.1 Introduction

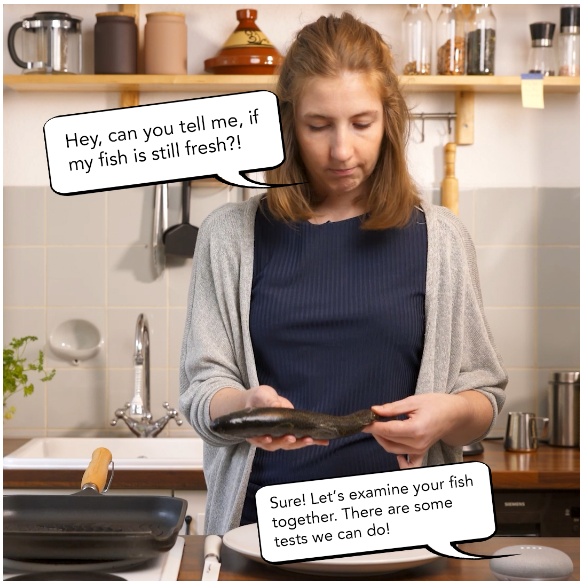

Hey, can you tell me, if my fish is still fresh?!
Sure! Let's examine your fish together. There are some tests we can do!

**Figure 4**
Asking Fischer Fritz how to assess fish freshness, own representation.

Food Waste is acknowledged as one of the major barriers to sustainable food systems in terms of environmental impact, food safety, as well as distribution in a world with a growing population [209, 242, 101, 64]. In the EU alone, nearly 88 million tonnes of food are wasted per year. Private households contribute to a majority (53%) of this food waste [305]. One reason for food waste, as pointed out by Hebrok et al. [134], are the insecurities of consumers regarding interpreting date labels and assessing the state of food. To make the right assumptions and decisions, consumers require “food literacy”, which can be understood as the competent application of (embodied) food knowledge [330]. Consequently, the practical experience of sensory-based interaction with the world results in “trusting your guts”, that goes beyond simply acknowledging institutionalized forms of knowledge, for instance formally written down rules in cooking books [326, 134]. As modern consumers increasingly lack food literacy, the problem of decision-making regarding food safety is challenging and leads consumers to throw more food away than would be necessary from a safety perspective [134, 115, 109].

While kitchens have become increasingly smart and equipped with a variety of appliances from smart ovens that clean themselves to everyday helpers like the Thermomix [311], they are not capable to prevent food waste. Some HCI approaches [135] like bin or fridge cam[318, 65, 2] attempt to create awareness of the household’s food waste behavior, but do not yet address the source of insecurities of food safety. Besides, research in Human-Food Interaction (HFI) indicates that automation-driven technology might even compromise the rich and embodied interaction with food, thus potentially further impeding food relationship building [3, 133].

To address this issue, we propose an approach that utilises human-agent collaboration [357] to enhance embodied knowledge as “competence-to-act” [108] and to promote sustainable and conscious food resource handling. Intelligent Personal Assistants (IPA) resoectively conversational agents gained popularity in recent years as commercially available Voice Assistants like Alexa or Google Assistant and allow for ambient interaction at home without too much attention directed to the device [120, 357, 253, 114, 258, 145, 334, 12]. Even though they are still limited in terms of their skills, technology as such provides interesting potentials for empowering human action by providing context-dependent cues and instructions [334]. By studying both humans and IPAs in collaborative action and decision-making, we want to explore how humans perceive the agency and the role of the IPA with its qualities and limitations to support the sharing and application of embodied knowledge for food waste prevention. Furthermore, we attempt to derive implications for the design of domestic human-machine co-performance [177, 163].

Our design case study [362] follows a user-centered design approach that is based on the actual food (waste) practices of households with the aim to support and enhance their food literacy. Therefore, we conducted contextual inquiries in 15 households and interviewed six experts about their approach to assessing food quality. We have chosen fish as an application domain, as this is a particularly sensitive food that comes with the most insecurities for consumers. Based on the preliminary implications of our formative studies, we developed and implemented a voice assistant called ’Fischer Fritz’ that aims at supporting users in applying sensory-based embodied knowledge to assess the quality and state of fresh fish (Fig. 4). Finally, we created a scenario-based video-prototype to evaluate the experience and approach beyond usability and detailed functions. We evaluated the potential to teach and negotiate embodied knowledge and collaborative decision-making towards food waste reduction with experienced consumers, allowing us to learn about how to further improve our design for implementation in common households.

Our research highlights how AI agents can be designed for supporting situated learning and application of embodied knowledge. By means of Research Through Design [236, 104, 31], we contribute the design of a prototype providing support for assessing the edibility of food to potentially reduce food waste practices in households. Our study further extends related work on HFI by complementing automation approaches with expertise building to reduce insecurity in food quality and edibility assessment without compromising the human-food relationship. Finally, our case study demonstrates how future domestic co-performance might contribute to empowering humans by increasing their competence to sense, think, and act on food (waste), ultimately contributing towards more sustainable food practices.

## 7.2 Related Work

### 7.2.1 From Lack of Embodied Knowledge to Food Waste

Food waste results from various factors including demographics, lack of routine and planning, inappropriate storage, aversion to leftovers, and even a lack of cooking know-how [280]. All that factors are entangled within the complex nexus of household practices [101]. Freshness of food is a multi-dimensional and cultural-historically shaped concept, where a specific meaning depends on the background of the person [44] as well as their everyday life context [245]. Since childhood, eating habits have developed into one of the most stable habits with every eating experience and sensory perception. They remain non-reflected for a while and are shaped by the social environment, upbringing, and education [213, 336].

However, one reason that should not be underestimated, as it contributes to a significant share of food waste, goes along with the understanding of date-labeling and food perception [355, 326, 280]. Especially concerning refrigerated products such as fish or meat, consumers are uncertain about consumption and tend to dispose of food [326]. From a practice-theoretical perspective, Hebrok et al. [134] argue to decrease insecurity when consumers asses edibility between institutionalized knowledge, embodied knowledge, and sensorial perceptions. Gherardi and Nicolini [108, 109] define knowledge as a “competence-to-act” by negotiating the “meaning of words, actions, situations, and material artifacts” with people and resulting in “practical accomplishment”. Thereby, applying knowledge becomes observable. Whereas, institutionalized knowledge is theoretical knowledge, e.g., labels such as “best-before date” or explicit rules, e.g., for the storage of food, written by authorities or non-governmental organizations [115, 109]. Embodied knowledge, on the other hand, is built up through prior experiences, e.g., the sensory evaluation of tasting, smelling, seeing, or touching food [115, 212]. Yet, especially knowledge on safety is frequently formalized and institutionalized, but "does not produce safety by itself, but only when it is put to work by situated actors in situated work practices and in local interpretations of its meaning and constraints." [109]. Although embodied knowledge is formalized into rules to some extent, it still has to be appropriately recognized or applied by its users. Due to little experience in sensory evaluation and trust in formalized regulations, consumers tend to prefer institutional knowledge, especially the best-before date, which leads to unnecessary food waste [326, 134].

Here, Alan Warde [343] argues that, due to the lack of embodied knowledge, consumers cannot perform the practice of assessing food and remain in a reflective state of mind. Interventions should therefore focus on reconnecting consumers and food and promote bodily experiences and memories related to food properties and conditions [338, 311, 319]. Meanwhile, consumers might train their embodied sensors, evolve trust in the understanding of their sensory perception, and, hence, obtain the competence to act in any given situation. The endless repetitions in the same place and the same time contribute to establish strong habits grounded in purposeful action to waste less food [354]. During learning, contextual information and rules should be embedded in the situation to guide the practical application of knowledge [134, 109] before everything is internalized as practice [354, 212]. With somebody acknowledging this participation as competence and people reproducing the results repeatedly from a first-hand experience, practices might be established over time [108]. This is called purposive learning as a form of social learning and apprenticeship that focuses on active bodily and mindful participation by observing rules and procedures, accompanied by guidance and feedback [354].

### 7.2.2 Human-Food (Waste)-Interaction

Due to its high environmental significance, preventing food waste has been a prevalent topic in sustainable HCI research for over a decade [135, 160]. Similar to food waste research in general [256], HCI approaches are dominated by behavioristic and persuasive approaches [135]. Some research [318, 65, 2], uses bin cams to post images of waste on the Facebook account to promote social comparison and pressure. These design interventions lead to increased awareness and interest in improving personal food waste disposal skills. Furthermore, Lim et al. [192] use direct feedback on discarded food to stimulate self-reflection and improve food planning. Other studies focus more strongly on supporting planning and self-reflection behaviors, like, for example, fridge cams [101, 93]. Those devices record all interactions regarding the fridge and allow access to the contents of the refrigerator ubiquitously. Moreover, research that focused on improving fridge management [192] and giving shelf life reminders [360] observed more awareness and hints for a reduction in food waste. Still, managing the inventory by hand and tracking consumed goods are tedious tasks that might not solve the problem in the long term [99].

In the light of HFI research, Bertran et al. [3] argue for a critical reflection on the agency of technology to not compromise the rich and embodied interaction with food. To bridge the gap between awareness and action, practice-oriented researchers such as Ganglbauer et al. [101] call for the promotion of *"specific practices in which "food is done" to promote more sustainable in-the-moment choices"*, which — against the background of food waste literature [134] — asks for more engagement in the embodied moments of deciding whether to prepare and consume food.

From a more celebratory perspective [118], HFI engages in the embodiment of the sensorial perception of food [234, 235]. This branch of research focused on the design of gustatory interfaces that simulate taste [259, 226, 328], touch [24, 127] and smell [159]. Still, what we can learn is the embodied reaction of users to these impressions that comes with emotions and full-body reactions [234]. An emerging theme is the need to engage these experiences in interaction with real food. For example, Vannuci et al.[327] call for more design towards cooking as a craft where technology enhances the cooks' agency in "touching, smelling, tasting, listening, speaking and enacting choreographies with the materials at hand". Similarly, Hassenzahl et al. [133] argue that rather than enhancing the technology's agency, we should engage users and their senses, let them experience their competences, and connect them with food.

Food waste research identified the lack of (embodied) knowledge of the sensorial characteristics of food as the main cause of food waste [134], yet, this discourse is currently missing in the sustainable HCI literature. However, sustainable HCI just began engaging in handmaking and sensorics [160, 135] and "encourages hands-on learning about food materials and nurture commonsense food knowledge instead of prioritizing automation and standardization" [79].

### 7.2.3 Co-Performing Conversational Agents

To enable humans to use their senses and enact in choreography with food, a user interface is required that allows for interaction with both: the food and the device at the same time. Here, IPAs respectively, conversational agents offer promising solutions as they do not need visual contact nor occupy the touch sense during the interaction. And indeed, commercial voice agents, like Siri, Alexa, or Google Assist, are increasingly pervasive in kitchens to assist in various situations. Thus far, they are used for short and trivial actions, such as setting a timer [120, 289, 6], but bear the potential to support the human with complex tasks and decisions [120, 357, 253, 114, 258, 145, 334]. Furthermore, research shows promising results to use conversational agents as learning environment or companions [102, 200, 71, 137]. However, as Hobert & Meyer von Wolff based on their literature review conclude, generalization, e.g., design knowledge and a thorough understanding of the design process is missing to contribute to future design of valuable learning environments.

Nevertheless, designing more complex tasks for IPA is difficult. First, the attempt to mimic human-like capabilities leads to high expectations in the intelligence of the assistant [197, 53, 275]. Up to now, these are mostly not met and leave users frustrated with the limited relationship to the agent [53]. A similar phenomenon is observed for social robots [273, 74, 325, 285, 199] where the anthropomorphic or human-like design implies a social presence which they do not live up to in direct comparison to humans. Here, distinct roles with an accordingly defined skill set [340, 199], with speech as a functional, embodied communication feature [197] might be a better paradigm.

Second, despite the opportunities of conversational agents to allow for human agency, they often miss the chance to engage the human directed in action in the decisive moment of the situation [268, 53]. Most dialogues are designed for simple command-responding tasks where the human commands the agent to execute [289] or the technology design in general rather focuses on automation and eliminating human decision-making at all [210, 117, 5]. Consequently, the design space for collaborative decision-making and co-performance remains secondary.

The notion of co-performance addresses both issues by exploring a useful distribution of capabilities and responsibilities [177, 110, 111, 163]. The authors [177] argue that an artifact should be considered as an active contributor to practice and designed to have the autonomy to learn and act next to humans. For example, they studied domestic heating practices and their evolution of the involved artifact, regarding capabilities, responsibilities, and roles in collaboration with the human from the fireplace to thermostat. While in the past the judgment to heat the fire was assigned to the human, nowadays the thermostat has the agency to decide about the temperature. Still, the human with his or her senses might experience temperature differently. Depending on the situation and embodied knowledge of temperature regulation, the human might overrule and negotiate the decision of the artifact. In this sense, the performance and decision-making of an artifact should be discussed under technological terms in the realm of possible sensing, interpretations and actions that are differently embodied than by humans. Kim and Lim [163] discerned in their study on human and agent co-performance influencing factors like the human's mental model towards the agent, considering a learning period to build trust in decision-making and that applying more human-likeness does not contribute automatically to more acceptance and rapport-building. The "artificial performers should be considered as a category in their own right and not as (poor) imitations of humans ones." [177]. Instead, we should focus on the design of an appropriate process of collaborative decision-making exploiting each one's capabilities, ecspacially in situations of uncertaintity [197, 190]. Form an embodied cognitive science view, according to the Sense-Think-Act cycle of Pfeifer & Scheier [250], intelligent machines have first to sense and then to compute before they act situated. Situated means "if it acquires information about its environment only through its sensors in interaction with the environment" [250]. Yet, sensors of machines might not capture the situation in full multi-sensory as humans do. In the opposite, humans often sense and perform certain practices simultanously with less deliberation involved. Yet, in the case of food assessment, for instance, they have to actively reflect on their intentions and multi-sensory perceptions first. Therefore, our research question focuses on how voice assistants may contribute to negotiating (embodied) knowledge with humans and potentially to preventing food waste.

## 7.3 Pre-Study: Edibility between Shelf Life, Rules-of-Thumb and Trusting One's Guts

Gaver [106], as well as Wulf [362], argue that the aim of a pre-study is to sensitize and inspire design. In this methodological tradition, we descriptively present our pre-study. The objective of the empirical pre-study was two-fold. First, we aimed to understand the current practices of consumers, how they examine the edibility of food, which knowledge and skills they apply, and how they negotiate institutionalized and embodied knowledge. With the main problems of consumers well covered by previous research (section 2.1), we summarized relevant design insights to our specific case. To further understand the assessment of fish and how to explain the procedure to an apprentice, we interviewed six experts. Although food waste is present in all food groups, especially in dairy products as well as vegetables, fish is a particularly sensitive example that is subject to many uncertainties of consumers. Hence, it is exceptionally challenging and risky because a majority of consumers lack knowledge and experience with this product.

For the first part, we conducted a qualitative study with 15 consumers (C1-C15), using semi-structured interviews and contextual inquiry in their kitchens. We took photos of the inside of their refrigerators and asked them to explain and show their everyday food handling to further understand the material context of different performances of storing food and assessing freshness of food. The participants have been advised not to prepare for the interviews because we wanted to observe their actual practices, e.g., maintaining freshness of products that are overdue. All participants testified that the inside of their refrigerator has not been altered for this interview.

The participants were recruited through opportunistic sampling within the author's extended social network. The sample varies in its socio-demographical characteristics with 11 female and 4 male participants, aged between 18-88 years, but having the main responsibility of household management and food practices, as can be seen in Table 4. Furthermore, it ranges from younger inexperienced consumers to family parents with a lot of cooking experience. Due to this diversity, we were able to identify a variety of food practices. For the second part, we conducted semi-structured interviews with six experts (E1-E6), including a university teacher on food safety, a cooking teacher, fish traders (supervising apprentices), and a chef. First, we asked them to explain their assessment procedure to a trout that we had brought with us. Next, we followed a semi-structured interview guideline to understand their explanatory approach, recommendations for consumers, and risks. All interviews were transcribed and analyzed in MAXQDA [2] following the inductive approach of thematic analysis [62]. Accordingly, the answers of the participants were coded and clustered by two

[2] www.maxqda.com

researchers independently. We discussed the codes among the authors and refined those in a second round to derive the themes of our analysis.

**Table 4**
Overview of Contextual Inquiry Participants

| ID | Age | Gender | Profession | Household |
|---|---|---|---|---|
| C1 | 52 | f | Nurse | Family (2) |
| C2 | 27 | m | Student | Shared flat (3) |
| C3 | 24 | f | Student | Shared flat (2) |
| C4 | 54 | f | Bank employee | Family (4) |
| C5 | 25 | f | Student | Shared flat (2) |
| C6 | 20 | f | Student | Family (4) |
| C7 | 48 | f | Housewife | Family (4) |
| C8 | 57 | f | Lawyer | Family (5) |
| C9 | 22 | m | Student | Family (4) |
| C10 | 59 | m | Employee | Alone |
| C11 | 22 | m | Police Officer | Alone |
| C12 | 51 | f | Accountant | Partner |
| C13 | 22 | f | Student | Shared flat (3) |
| C14 | 53 | f | Accountant | Family (4) |
| C15 | 56 | f | HR Manager | Family (4) |

### 7.3.1 Consumers Approach to Assessing Food

We found varying strategies for assessing the edibility of food for different products. For packaged food, canned food or jam (C10, C13, and C14), milk (8 of 15), and cheese products as well as meat and sausage products (6 of 15), the shelf life is used as an initial indicator. For some consumers, shelf life is not critical for their consumption decisions, e.g., for meat (C1, C5, and C14) and dairy (C1, C2). Some consumers would even buy and consume those products with an expired date if they can consume it the same day. For others, shelf life varies between guiding and determinant when disposing of food.

Problems arise when participants no longer can recall the product opening, purchase date, or expiry date. This is resolved either by sensory evaluation of the products, or for some by estimating the time (C4, C12, C14, and C15). At this point, all participants declare their intentions and attempt to use their senses when examining the freshness or edibility of food before they prepare, eat, or dispose of it. For this purpose, they begin with a visual assessment, looking for signs of decay, e.g., mold or rot. This procedure is conducted for any product. Regarding milk and yogurt, the participants declare that they are impervious to shelf life if the consistency has not changed and no mold is visible. First, they smell the product and then eat or drink a small amount in the meaning of a "small spoon" (C1) or a "knife tip" (C12), which are harmless to health, to further decide on the product. The spoiled smell is described by C7, C11, and C14 as acidic and C15 would explicitly look at the milk to see if it "crumbles". However, in the case of fish, meat, and boiled eggs (C5, C11), participants expressed greater concern about food poisoning. This is why they act much more cautiously and some tend to throw the product away. Here, also the consistency of meat is checked for changes in color or “smeariness” (C10) and its smell (C11, C15). The eggs are, if possible to estimate the storage time, at least peeled and checked for optical and olfactory signals. Nonetheless, participants state that some qualities cannot be assessed by their senses. As C8, for example, explains, a salmonella infestation cannot be detected.

Interestingly, we could observe varying degrees of food literacy regarding age and household responsibility. Student consumers more often lacked consistent routines and competences to maintain the freshness of food and storage hygiene. Whereas older and experienced consumers had appropriate storage solutions like special tupperware but also more space such as a second freezer in the basement. All in all, the explanations of the consumers show them trying to triangulate between their bodily reaction to the food, rules such as the identification of “crumbliness” and institutionalized knowledge in the form of shelf life. Especially when one information source does not lead to a decision, uncertainty arises and different approaches are combined. C14, for instance, explains that the visual perception of meat on the verge of expiry must be perfect, and stressing the meat should not “leak”, referring to liquids inside the plastic container. To double-check edibility, she does an additional smell test. To obtain additional information, C15 also asks a person for their opinion on freshness and shelf life.

Edibility turned out to be a complex, culturally related construct that is also shaped by individual horizons of experience. This experience is usually described in years of experience or gained through cooking with parents. Concerning this, the perception of edibility differs. In this context, several participants also talk about freshness, which seems to be used as closely related to edibility. Nine participants describe freshness as rather “harvest fresh”. Some of them refine it as “from field on the table or directly to the stomach” (C12) and as “ultimate freshness” (C11). Besides, ‘freshly harvested’ also means that the product is just ripe (C4, C8, C10) as it “falls from the tree” (C10). Furthermore, participants used nonsensory characteristics to define edibility. Seven of the participants associate it with healthy-to-eat and safe food. For one participant, however, food safety is nowadays even of secondary importance to environmental considerations:

> *“Sterility is the wish that things are packed that not everyone has touched. Today it is rather that I take the unpacked goods because I would like to support environmental thoughts.”* –C15

Freezing of food to preserve edibility is, however, controversial. C3 and C12 regard shock-frozen fruits and vegetables as vitamin-rich as freshly harvested products. In contrast, five participants judge frozen products in general and, more precisely, defrosted bread or ready meals as not fresh.

### 7.3.2 Teaching Embodied Knowledge

As our contextual inquiry confirmed, perception of freshness as well as assessment procedures differed between the households with meat and fish as particularly sensitive cases.

Against this background, we wanted to focus in our design on this food item as the model case. Asking the participating experts to explain how a fish should correctly be assessed in terms of its edibility, it quickly became apparent that a multi-sensory approach is needed. This approach includes the senses of sight, smell, and touch that are applied to various characteristics of a fish. Taste, however, is according to E1 only appropriate if the fish is processed in a salad or similar.

All experts agree on such a multi-sensory approach as different preprocessing steps might change certain qualities of the fish. For example, storing it on ice clouds the eyes, which is usually perceived as a sign of decay. However, also some tricks and attempts to deceive consumers were explained. For example, E2 examined a fish with the gills removed from the fish, which the expert calls as a trick to prevent the fish from bad smell and to cover up the non-freshness. Besides, fish can also be prepared with additives such as lime juice or slightly smoked, to enhance durability and hinder the freshness assessment.

In summary, the assessment procedure introduced by the experts includes the following test items. Still, not every expert uses every of those test items, as they usually just need a few checks to determine edibility. However, as explained, multiple tests might be needed if, e.g., the gills were removed.

- checking for the smell (either *neutral* of fishy)
- checking the flesh with pressure (either the dent stays or *not*)
- visually checking the eyes (cloudy or *clear*)
- visually and tactile checking the skin (*slimy and shiny* or dry and dull)
- scratching the scales with a knife (falling off or *not*)
- visually and tactile checking the fins (dry and frayed or in *wet and normal conditions*)
- visually checking the gills (*red and not slimy* or pale color)
- visually checking the inside (*light red* or thick/coagulated)
- visually checking the flesh (*normal color* or greenish/brownish)

Moreover, those rules do not have an explicit order, but some experts used the saying “the fish stinks from the head” (E3) to explain how they start with the gills and their smell. They continue with the eyes and go on with the other parts. Still, some of the items are considered as stronger and more obvious indicators for peak freshness like fire-red gills.

As this rule-of-thumb already indicates, assessing the freshness of a fish is similar to the approach of our interviewed unprofessional consumers, and closely related to experience and some roughly defined rules. Much knowledge is embodied, and over the years, the experts learned to understand their bodily reactions and feelings. For example, E5 said “Bad smell you know, my sense of smell will understand it. It's non-describable. It's kind of abstract". Nonetheless, they tried to articulate their knowledge as rules, for example, using analogies such as "fishy" or "seaweedy" for bad fish or "neutral" or "fresh sea breeze" for good fish. Similar articulations were found for visual characteristics, such as "bloody colored" or "rose". Here, they also often referred to a normal-looking fish that they had internalized over time. Quite difficult was the articulation of the tactile sense, which the experts indicated to, for example, the normal reaction of the fish skin to pressure, which is fish type-specific and must be learned with time. Still, they argued that a fast reaction of the skin to the pressure is a good sign. Moreover, they highlighted to show and explain the location of certain body parts, e.g., where to find the gills (E6).

Finally, the experts raised our awareness about the field of tension between sustainability and food safety. While some experts (2/6) were more relaxed to the danger of eating slightly decaying fish, others recommended being more cautious. In the worst case, the fish can be toxic, but still, they argue that in those cases everybody should show some natural bodily reaction. Furthermore, in cases of doubt, they recommend at least well cook the fish to prevent salmonella. In this respect, the shelf life was mentioned and that any fish, far from this date, should be disposed of. Otherwise, sensory assessment should only be used in doubt near the shelf life.

### 7.3.3 Preliminary Implications for Design

As the results show, consumers are motivated to use their senses, but often miss guidance on the procedure and interpretation support. Prior research [101, 134] already highlighted that more support for in the moment decisions is needed since lasting behavior change is challenging to achieve. Furthermore, our research points out that the meaning of freshness is affected by the dispersed moments of consumption practices [101]: During shopping, freshness is described as harvest-fresh and ripe, yet, descriptions change in the home context, where the focus is rather on the assessment of edibility. Hence, storing does not represent a negligible practice within the nexus of consumption practices [101, 3, 184], but is central to ensure freshness. It is a practice of keeping food as fresh and edible as possible, in need of competencies to assess food qualities by making use of multiple senses and food condition information [92]. Therefore, we should offer advice beyond the obvious visual indicators of decay and provide clear, quick-to-apply instructions that promote experiential learning and collaborative learning. We need to explain food safety regulations in context, and use descriptions that illustrate the gradual differences in food quality like, for example "sea-weedy". Further, our findings show that freshness is often described negatively as a deviation from expectations, how something must look, taste and feel [245]. Therefore, antonyms are used such as "not old", "not spoiled" or "if the salad is not withered" to define what is not fresh. This indicates the importance of verbalizing sensory impressions that contribute to a shared understanding, in particular when designing with speech. Furthermore, consumers have to train senses to trust their bodily reactions and develop personal rules-of-thumb. The freshest food offers the highest taste experience, but consumers need to taste first to know the best condition of the food. Moreover in everyday life, trade-offs between fresh and, therefore, healthy food and edibility cannot always be sufficiently avoided. Hence, the prototype needs to enable consumers to understand that although food can still be processed, it may require additional flavors to improve the taste. The design has to acknowledge the perceived severeness of varying health risks between food groups by the consumers to ensure sincerity and reliability. With fish and meat, consumers are very critical and cautious and tend to dispose of the food more quickly. Yet, the prototype should refrain from blaming consumers if they do it anyway. Instead, we need to carefully and patiently explain the instructions and assessment categories as transparent and comprehensible as possible.

## 7.4 Prototyping: Voice Agent

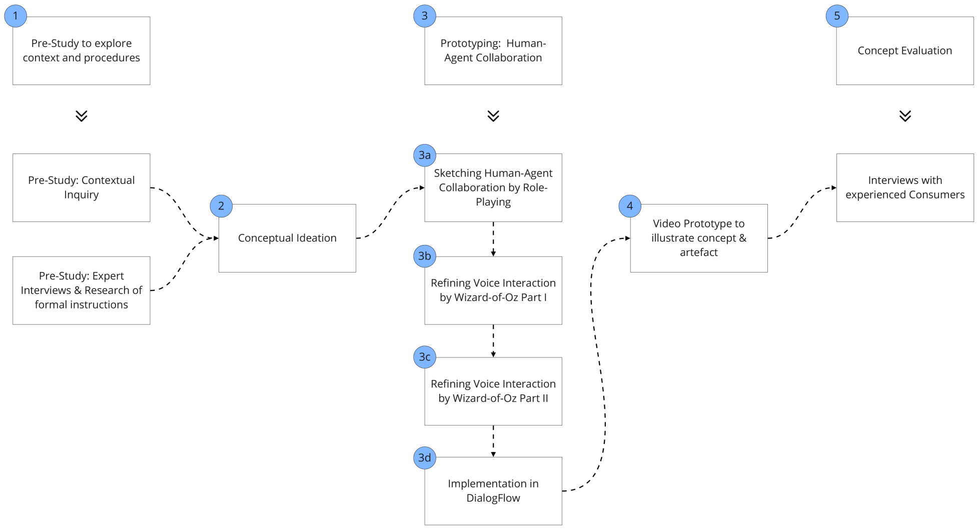

1
Pre-Study to explore context and procedures
Pre-Study: Contextual Inquiry
Pre-Study: Expert Interviews & Research of formal instructions
2
Conceptual Ideation
3
Prototyping: Human-Agent Collaboration
3a
Sketching Human-Agent Collaboration by Role-Playing
3b
Refining Voice Interaction by Wizard-of-Oz Part I
3c
Refining Voice Interaction by Wizard-of-Oz Part II
3d
Implementation in DialogFlow
4
Video Prototype to illustrate concept & artefact
5
Concept Evaluation
Interviews with experienced Consumers

**Figure 5**

Proceedings and single steps of the Design Case Study, own representation.

Our first expert interviews and observations verified the main criteria and sensing approach for quality assessment and helped to determine the best guidance order and provide additional reasoning for fish characteristics. To a further extent, we triangulated the preliminary implications with research-based guidelines on fish food safety [67]. In the next step, we carefully solidified the empirical data in a collaboration model to define both, the procedure of assessing freshness as well as the capabilities of the user and voice agent (Fig. 5, step 1). Based on the preliminary design implications and needs consumers have, we aimed to design the food assessment as a collaborative task (Fig. 5, step 2). As follows, the main concept was iteratively evolved using a combination of Role-Playing and Wizard-of-Oz sessions [116], as can be seen in Fig. 5 steps 3a to 3c. To investigate the procedure and elaborate dialog drafts, we began with Role-Playing in our team. In contrast to Wizard-of-Oz, Role-Playing allows to explore dialogs freely to collect possible directions and phrases. As a rigorous method to test the system's capabilities by the efficiency and sufficiency of utterances, we continued with seven scripted Wizard-of-Oz guidance sessions that restricted further use of "common-sense" to empathize with the user [38]. At this stage, the agent already had some structured guideline with questions and answers to adhere but still the wizard was able use some common-sense to prolong the dialog to a successful ending. After refining the dialog paths we conducted a second round of Wizard-of-Oz. This time we used the telephone to reduce a potential social presence of the agent and did not deviate from the script. This allowed us to rework error handling and fallbacks by experiencing dead-end conversation cues. Our sample was between 20 and 30 years old and unfamiliar with fish assessments. At this prototyping stage, the limits to a design agency and coaching became aware. The interviewer, in the role of a voice agent, used the list of attributes that indicated the status of freshness. The potential users had a photo of the fish for greater immersion in the situation. The drew upon past encounters with fish and imagined different states of sensory impressions. We renounced the use of fish in the prototyping phase to avoid food waste. Finally, we implemented our dialog tree [3] in Google Dialogflow (Fig. 5, step 3d) and tested all paths (Fig. 6) within our team. Afterwards, we captured the interaction between the user and the agent as one "happy path" in a video (Fig. 5, step 4), to use this video-prototype to illustrate and evaluate the conceptual design of the artifact (Fig. 5, step 5).

### 7.4.1 Sketching Human-Agent Collaboration

The first draft of our concept was based on the main assessment criteria from our prior research and food quality experts. Furthermore, we used the observations and suggestions by the experts to prioritize the chronological order of information, so that users get reliable results with a minimum number of questions. Therefore, we visualized the potential paths and outcomes in a decision tree and specified the most critical characteristics to be asked first, as can be seen in Fig. 6. Assessments like gills, smell, and color of fish flesh are primary and mandatory aspects, whereas eyes, scales, and fins are additional determinants to indicate the condition of the fish quality. Nonetheless, the ambiguous interim results of the fish condition will need more checks for a final decision. We designed transparent step-by-step explanations to allow users to trace the decision path from beginning to end, e.g. *"Okay, so your fish has no gills. Then let us skip the gills and start with the fish inside test. Let us now open the fish, so that we can see the abdominal cavity. Is the fish meat bright and more to the whitish, pale, or pinkish or is it more to brownish yellowish or greenish?"*. Thereby, the agent encourages the human to interact with the food product and teaches to interpret the sensory impressions correctly to come to their own, resp. the same conclusion, as for example seen in Fig. 7. The human constantly describes and answers the agent to determine, collaboratively, and successfully, whether the fish is still edible without risking health. During the co-performance of the assessment, the users shall not feel patronized, but self-confident and reassured by the

[3] We developed and implemented a German version of the dialog.

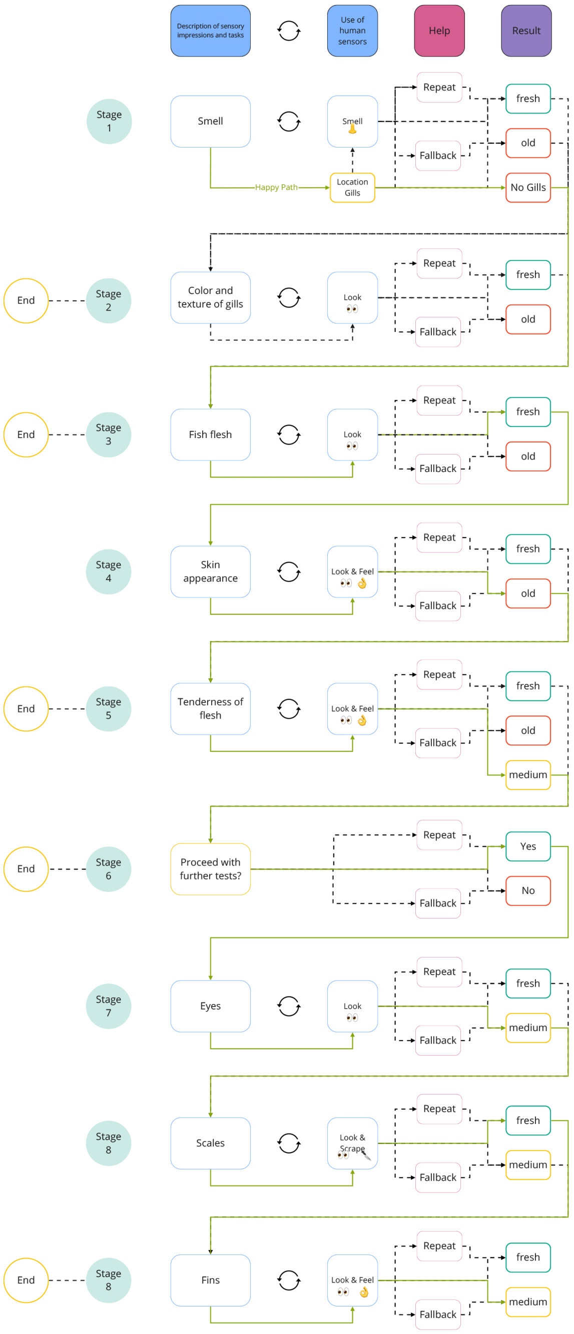


**Figure 6**

Representing all possible Dialog paths for human-agent co-performance. In total, 8 indicators to check freshness with 5 possible conversation endings on behalf of the user. The video prototype showcases the solid line from Stage 1 to Stage 9.

collaboration to trust their own senses: *"That is good. The body of a fresh fish is firm and when pressed it should bounce back. The fish is fresh enough to be prepared with heat but* *the flavour might not be the best as the skin is not at peak freshness. Would you like to further have more detailed info about the freshness of your fish?"*. Yet, the agent has to react patiently to possible misunderstandings or indecisiveness by users. Finally, the voice agent emphasizes that the responsibility for further actions lies with the user.

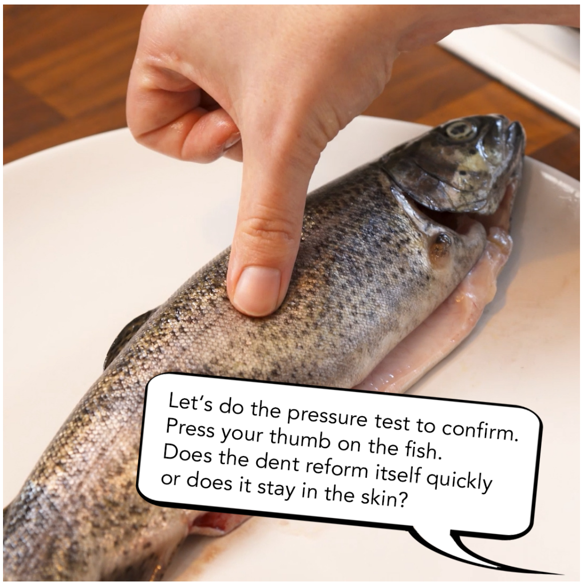

Let's do the pressure test to confirm. Press your thumb on the fish. Does the dent reform itself quickly or does it stay in the skin?

**Figure 7**
Instructions to perform pressure test for freshness by Fischer Fritz, own representation.

### 7.4.2 Refining Voice Interaction

In the following Wizard-of-Oz sessions, we explored the conversation with different degrees of role-play restrictions and freedom to simulate the intelligence of the system, as shown in figure 6 step 3a to 3c. Meanwhile, we noted possible dialog sequences, unexpected edge cases, missing fallbacks, and collected a variety of utterances to refine the dialog. Besides the right keyword use, edge cases include remaining challenges to explain sufficiently the position of the gills, the right amount of pressure on the skin or verbalize possible olfactory impressions. As ambiguous descriptions lead to misunderstandings, we implemented non-standardized fallbacks to catch edge-cases and to sound more personalized. Moreover, repetitions help to ensure a shared understanding of the progress and indicate active listening, as can be seen in Fig. 6.

Conversational guidance is based on proactive questioning and proposing distinct adjectives to simplify decision-making. Hence, the Voice agent is responsible to perpetuate the dialog and depends on users to answer. We deliberately reviewed all utterances and refined wording and sentences. Thereby, we decided to use explicit adjectives to provide users with clear answers to use. Some of our participants during prototyping find it hard to describe their sensory impressions in their own words. This results also in an advantage for the interaction, since potential dialog errors and fallbacks are reduced to a minimum. The trade-off is a less free conversation for the human, yet better than leading questions on a yes-or-no basis as criticized in our Wizard-of-Oz sessions. Furthermore, distinct opportunities to exit the dialog increase the satisfaction of an accomplished task. Either the agent ends the dialog by reaching a decision quickly or users are convinced to have enough information to skip some or all further assessment steps. Some test users mentioned that they liked the provided additional or more detailed information, but would prefer to ask actively for it. Further, for transparency reasons and to show trustworthiness, the final suggestion of whether to consume the fish or not is carefully verbalized and communicated: *"Fins and scales are in great condition but the eyes make it appear a little bit less fresh. The fish is good but please also rely on your senses to not risk your help."* To emphasize an inclusive understanding of performance and create a team experience, we used utterances like *"Let's perform a few tests."*.

## 7.5 Prototype Evaluation by Experienced Consumers

The main goal of our evaluation was to explore attitudes towards the usefulness of voice assistants in the prevention of food waste, their potential and limits to convey embodied knowledge, and decision-making in collaboration with our voice agent Fischer Fritz. The prototype is not exclusively designed for cooking novices, but aims to support where guidance is needed. We used video-prototyping as a common method in HCI to focus on the concept evaluation of novel artifacts, as proposed by Diefenbach and Hassenzahl [77] allowing to observe several experience levels like interaction, functionalities, and emotions at the same time. The attention is directed rather to the embedded everyday experience without distracting users with usability problems or immature technology aspects [77, 302]. This method is also suited for Human-Agent-Interaction [315, 141]. In light of our contribution, this work goes beyond a usability evaluation and discusses design implications to improve sensing, thinking, and acting in co-performance as immediate guidance in the situation of challenging indecisiveness based on a novel artifact. As solving usability issues was already in scope of the iterative technology design, the evaluation reflects on the opportunities of the design to promote appreciation of food and preventing unnecessary food waste.

The video-prototype takes 4:05 minutes and shows a typical scene where a consumer picks a fish from the fridge and doubts its edibility. In the next step, Fischer Fritz is approached for support. In the following, the user and the agent exchange information about the fish characteristics and interpretation of the indicators to come to a useful conclusion. To immerse the viewers of the video, close-ups of the fish help to build their own impressions except the smell. The interactive guidance represents one possible assessment combination out of eight combinations in total (see "Happy Path" in Fig. 6 from the original dialog. Although we wanted to display the consideration of all available fish characteristics and sensory impressions. In this take, some of the fish characteristics are ambigous in perception, which leads to the most insecure scenario of all available outcomes by using the prototype. Our aim was to confront the participants with a remaining risk to provoke insightful discussions about trust in their senses, the voice agent, and prior knowledge as well as their attitude towards technology in general.

Afterward, we interviewed 15 consumers with a varying range of food experience across Germany, following a semi-structured interview guideline that roughly covered the topics of their perception of co-performance and its potential. In particular, we asked about the meaning and communication of (embodied) knowledge regarding food quality and food value, the risk and control in the process of decision-making, distribution of roles and capabilities, and impact of user empowerment. We aimed for a sample that is well suited to assess the role of the agent and the challenges of teaching rule-based and embodied knowledge, as can be seen in Table 5. The participating consumers (P1-P15) were recruited from contacts from prior studies and the extended social network of the authors. With aiming to collect a variety of perceptions and opinions on the prototype and concept to encourage more embodied interaction with food, we chose consumers with different experience levels. Those ranged from highly experienced home and family cooks to professionals who blog about food or who were trained in gastronomy and give cooking classes. We have a tendency of more food experienced consumers, as they interact with inexperienced consumers regularly and understand their struggles in a more condensed manner. All participants were between 26 to 80 years old. Moreover, as our sample is familiar with the properties of fish, video prototyping does not limit the evaluation due to less sensorial experience, rather allows to center the focus on the verbalization and communication of knowledge. From a more pragmatic stance, we moreover, did not want to risk any food safety issue in a real-world trial or unnecessarily wasted fish (that we would have to let decay on purpose) in a laboratory setting.

**Table 5**
Overview of Prototype Evaluation Participants

| ID | Age | Gender | Job | Relation to Food |
|---|---|---|---|---|
| P1 | 36 | m | Research Assistant | Food Blogger |
| P2 | 35 | f | Teacher | Family, Vegetarian |
| P3 | 35 | m | Chef | Chef |
| P4 | 54 | m | Sous-Chef | Gastronome, Cooking Courses |
| P5 | 40 | m | Project Manager | Food Blogger |
| P6 | 41 | f | Journalist | Food Blogger |
| P7 | 52 | m | Product Tester | Marketing for Cookware |
| P8 | 80 | f | Pensioner | Family |
| P9 | 39 | m | Chef | Restaurant |
| P10 | 52 | m | IT Specialist | Cooking Club |
| P11 | 34 | f | Freelancer | Healthy Food Blogger |
| P12 | 26 | f | Media | Study of Nutritional Sciences |
| P13 | 33 | m | Master Butcher | Food Blogger |
| P14 | 36 | m | Media Designer | Food Blogger |
| P15 | 41 | f | Chef | Book Author |

All interviews were conducted using remote conference calls and sharing a private video link during the session. Afterward, they were transcribed verbatim and thematically analyzed in MAXQDA. We followed the thematic analysis procedure as outlined by Clarke et al. [62].

Two researchers coded the themes independently and discussed towards agreement on the themes for further refinement. The final themes are represented in the headings of the evaluation results. Furthermore, we translated the quotes from German into English.

### 7.5.1 Trust & Autonomy in Decision-Making

All participants are pleasantly surprised about the guidance by Fischer Fritz and the interactive design of the provided approach itself. They recognized their methods and explanations, similar to the way they teach knowledge to apprentices in the workplace (P5) or participants in cooking courses (P4). P11 emphasizes that inexperienced consumers could get a new perspective through the voice assistant and regain more confidence in their own senses. Some of the participants also praise the additional information, descriptions of the possible sensory perceptions, and explanations that are given for the respective fish characteristics.

> *The pressure test is very important. Sensational! Yes, you have addressed everything, everything important. There are of course many fish products that no longer contain gills. That can have many reasons, but it is usually said that gills are also decisive and must be bright red.* –P4

Overall, most of the participants (12/14) see an opportunity to reduce food waste and estimate the risk to make mistakes as low due to the distribution of tasks and capabilities. They (7/14) also value the availability of the technology at home and the easy access to information by Fischer Fritz. Even though some (5/14) of them express a certain distrust towards voice assistants, they confirm the potential support and comprehensible advice.

> *No, he [the voice assistant] is very clear and explicit, but I wouldn't say patronizing, it's just that he, and oneself, wants to make sure that everything is in order and properly inspected. If it is then later said 'yeah, I still got diarrhea' because the food was spoiled, then people say subsequently, 'he didn't tell me that I had to check it'.*–P13

Again, all generally appreciate the additional reassurance and note that, in their opinion, potential paternalism only arises when Fischer Fritz confronts users who have a higher level of knowledge than the agent himself (P8, P13, P14) or, for example, when new insights contradict their intuition (P2). Otherwise, they might blame the assistant and deny any responsibility. Moreover, users need to actively seek support when using Fischer Fritz (P8, P11), and they are prompted to make their own decision based on agreement with the results (P12, P13, P14).

> *The human is still [in control], and everyone should use his or her own mind or willing. Whether to eat it or not, he can decide for himself how he likes.*

> *Therefore, finally, I see the control still with the human and, so to say, the device only in such a way as a control body.*–P12

Moreover, participants (8/14) positively highlighted the structured and step by step guidance and sensory checks (P14) aligned with the actions of the user (P8, P14, P2, P10) without information overload. The descriptions help to check and classify the sensory impressions as well as to look at features that otherwise would not have been considered at all. Hence, the human takes an active role in quality control and retains autonomy in his decision making.

> *They complement each other. I think the machine has the knowledge and the human simply has the senses, which he has to provide.* –P14

### 7.5.2 Co-Performing Food Assessment

As mentioned by the participants (7/14) before, complementing the human, our voice assistant can eliminate the last uncertainty and contribute to autonomy in decision-making. For P9, personal control goes beyond his diet. Taking self-responsibility and self-care further lead to decisions for a sustainable environment, since interdependencies determine how we live together. This attitude implies a decision for conscious handling of one's own life and food.

> *I think if you eat a healthy diet, you tend to be more conscious of many things that concern you and also the environment. And therefore I would say, most humans I know, who eat very healthy, also pay attention to waste less food.*–P11

Besides, four participants considered using Fischer Fritz to check other foods. The value to save animal and plant products is compared to the effort required to use the voice assistant, and, hence, the probability of its use.

> *With this system, I think it would definitely be possible to avoid [food waste]. I could just imagine other examples that could be a bit more successful. Like potatoes or fruit and vegetables, simply where it is not that critical. The question is which foods should be prevented. Of course, high-quality foods such as fish and meat (...). All the dairy products, for example where you can still eat yogurt after 3 months. The fact that it is thrown away quickly. That the things that can simply be subject to longer storage, are also more likely to be thrown away like those that have a short lifetime anyway. (...) This is maybe with a yogurt that you have 2 weeks in the fridge and then after the 3rd week or a week past the expiration date you just don't know if you can eat it or not. This case is simply more relevant. The question is, whether in the case of a yogurt one would bother so long asking - answering, because it is also only a 15 cent product, and a fish*

> *may have cost 15 euros after all, that somebody is perhaps more likely to do that.*–P14

For the efficiency of information retrieval, most of the participants (10/14) compared the voice assistant with their usual Google search. Some of them (2/14) conclude that it would be faster to just read over the information quickly, whereas the majority emphasizes the situated learning and accompanied embodied experience. Some also add that specific information like why the eyes cloud are probably not found at the first search online (P2, P3, P4). Nonetheless, time-conscious participants (3/14) suggest having a quick overview of the total of fish characteristics at the beginning of the dialog. Thus, they can get the first impression of a high-quality condition of the fish. Further, P12 notes that there is always the possibility to skip some parts of the coaching by voice command "Further".

> *Quite well, because the written form is just, I think if you want to have a quick look, whether it is still good or whether it is already spoiled. Then it is so cumbersome to enter it somewhere, then just look for something or look it up somewhere. Same with the video. I had to find something first and I want to know it directly. And that is why you simply talk into the room, tap your cell phone, the voice assistant turns on. For me, this is one of the easiest ways to do it, instead of having to look for something somewhere and read it.*–P12

Furthermore, they reflect on the modality of speech and its appropriateness to convey knowledge. P3 and P8 note that it might be difficult to teach someone how hard to press on the skin. Although many of the participants (10/14) mention they use visual media like Youtube videos and TV shows, they rather watch it for inspiration than step-by-step guidance.

> *But what is shown today, I can only say: forget it. I really do not watch any more. (...) Surely anyone can grate or chop carrots. I do not necessarily have to show it on TV every time carrots are needed somewhere, I do not have to show it every time. All I need to say is, 'I think carrots belong in there or something.'*–P8

Moreover, some argue that pictures to compare the same types of fish (P3) or videos of embodied movements (P8) would support learning significantly. In contrast, P7 expresses his concern, how pictures may contribute to more insecurity by prescribing implications that are not appropriate for the fish at hand. Furthermore, P9 elaborates on how book authors are capable of creating images using comparative examples and words only. He suggests further to update the dialog in this manner.

> *I think examples would still be important there, which can produce such images in the mind (...). For example, the case of 'what no longer serves'. And then also creating the smell for "what no longer goes" on the mind. If you then have such an old Harzer cheese in front of you, so the fish smells like an old Harzer, then*

> *you still have to know now, how does a Harzer smell, but having so 2-3 examples of what people might know how something smells.*–P9

P1, P2, and P7 weigh in that technical features like cameras, scanners, or sensors could offer a technological sensory reassurance for the assessment. At the same time, however, they claim that it would counteract easy access to the technology already available and would require further investment. However, most of the participants (9/14) denied additional sensors, because they see reconnecting to food and using the human senses for this purpose as the most valuable.

> *If it is just a camera, you hold it in front of the unit, but if the system itself can touch and feel, I could imagine that you put the product somewhere on it and that it is scanned, touched and sampled. And you certainly know completely detached from the knowledge and experience components that this product can be processed. So you just don't learn, you don't train, but you completely hand over everything.*–P1

In direct comparison with human-human interaction, the participants (4/14) notice differences as they miss some emotion and passion in the interaction describing it as too functional or informative only. Moreover, P2, P3, and P6 see emotions as a key aspect for cooking and food in general. On the other hand, everyone emphasizes the purpose of Fischer Fritz and its contribution.

> *The interaction between each other, if you were to ask me now, 'is the piece of meat still okay' and then I could explain directly "aha here and there and that's how you see it,“ take it in your hand, etc. that's just not given with the machines. The cooperation, the communication among each other is different. That just doesn't work with a machine. But apart from that, it's completely okay, because it's purely informative - you want to know something from the machine, and that's why I think it works.*–P13

The majority of participants cannot agree on the role of our agent in the collaborative practice. Some (4/14) say 'assistant' is already a good choice because it provides useful advice and is informative. Other participants (6/14) think of it more caring and engaged.

> *I think of a mixture, I ask my mom how I do it when I'm cooking and really a kind of cooking teacher. Well, I don't think it's a kind of a true instructor. There is the issue using speech only, perhaps too imprecisely.*–P3

Concerning the voice interaction itself, P9, as an instructor himself, immediately felt strongly reminded of a training situation by the "tone of voice, by the way he spoke to the person". But to be able to speak of a "coach", in contrast to humans, the participants (8/14) miss traits like empathy, truly open questions, and spontaneous dialogues. Moreover, most of the participants (12/14), emphasized the explorative character of coaching, such as letting the users make mistakes and guide them to find their own solutions.

> *A coach helps, so I think what makes a coach is he helps you to develop or discover something yourself. He does not prescribe it but helps you to develop or discover the solution. He does not give you the solution, but he helps you to create it.*–P9

### 7.5.3 Embodied Human-Food(waste) Interaction

To increase the value of food and develop passion, the majority of participants (11/14) point out that people must engage with food and relearn the natural characteristics of food. Hence, some of the participants highlight that the interaction facilitates shifting the attention of the users to the food itself.

> *Yes, well, I just thought that it would be better now more practical than a book with my fish hands, or in the iPad, cell phone with my fish finger must search and the eyes are not for "seeing", etc. and that I also do not have to look anywhere, on a video, but that I can look at the fish all the time. So that I perceive auditorily, so to speak.*–P6

Still, the evaluation of food quality or safety without any experience is a challenge. In general, freshness is according to P3 and P4 a stretchy term. P6 mentions insecurities in online requests to her regarding the use of two or one tablespoons in a recipe, that are most of the time not decisive. However, generally deciding on the right ingredients, differentiating between high quality and edibility as well as recognizing the little difference to improve the taste requires experience.

> *Not fresh anymore means you have to put a little more love into the product when cooking, so that it still tastes good afterward. But inedible and, people are afraid of diseases. You have to know how to avoid it.*–P14

Concerning the leftovers of a product, for example, potato peels can be baked to ashes in the oven and mixed into mashed potatoes to intensify flavor (P4). But such stimuli come often as external impulses and need to encourage users to try. However, according to the participants (10/14) the successful application of novel information leads frequently to new personal confidence and in the information itself.

> *But also that a lot of people don't know that they can also eat the stem of broccoli when they cut it into small pieces and cook it. (...) That many people simply*

> *don't know what they can use from the vegetable or plant. Cooking experience definitely plays into that. So I've read up on it, but from experience I've tried it and found it to be good. I would not have had the idea to eat the stem on my own, because you are used to eating only the florets. And that I have read and tried it somewhere. You have to take that step, yes.*–P12

At the same time, the direct use of information and the associated experience contributes to engagement and relationship building with different foods. Therefore, people have to learn quality control slowly step by step, for example, what a good or bad fish means (P4). Long-term information and experience will transform into knowledge and help to live independently from technological systems for the most.

> *I think the system is good, if the system is ultimately used to learn to be able to do without the system at some point. If you make yourself increasingly dependent on the system, you might not even know what you can eat sometime in 10 years. Therefore, I think it is a support to find back to your own senses.*–P1

In this respect, there could be even more self-reflection promoted. Therefore, participants (9/14) claim it needs frequent and situated opportunities for novel topics and actions. Even before the presentation of the prototype, the participating experts (10/14) agreed that experience is gained through experimentation and that people need to be sensitized or confronted with it over a long period of time, in the best case, in comparison to the last experience.

> *By encouraging and motivating him [the user] to reflect holistically. To relive the experience. To repeat it more often. However, in the end, it is enough to re-ask about the situation. To ask yourself again 'Okay, how slimy was the fish now compared to my last fish? Remembering that.'*–P9

Participants (4/14) note that the quality of teaching and training depends on the user type and the way of teaching by the assistant. Therefore, the voice assistant needs the ability to learn and remember personal information about users like allergies or last requests. This gives the chance to track progress and build on previously acquired knowledge (P9). Moreover, the perceived role and function of the assistant, whether as a strict instructor or a friendly family member or coach, might impact the learning effect (P4, P14).

> *I think such an Alexa can sound very smart-ass but maybe there is another way. And then it is pleasant again. Or if people just want to be more factual, or fast and effective, the learning types are very different, how someone understands something, whether you need more repetition or not, and if the tool can do that. If the tool can do that, then I think it is already a great opportunity.* –P10

## 7.6 Discussion & Implications

Food Waste was addressed by various HCI prototypes [135]. While current practice theoretical research [134, 101] highlights the importance of sustainable in-the-moment choices, with a special focus on food quality and safety as well as the value of food, prior HCI research primarily addressed food waste as a motivational issue [135]. As we used Research through Design [236, 104, 31], we contribute to a thorough understanding of the design process of interactive agents for a learning environment [137] and outline a potential co-performance by our conceptual design [177]. Accordingly, our design approach is accounting for those decisive moments that are, according to Hebrok et al. [134], entangled between embodied and institutionalized knowledge, e.g., labeled dates. Hence, we reflect with experienced consumers on the potential impact, trust, and responsibility, as well as the necessary artifact properties in the decision-making process contributing to sustainable food practices. Reviewing our voice agent and the respective design case study, we want to discuss our research along the Sense-Think-Act Cycle by Pfeifer & Scheier [250] as a guiding design principle. Usually this model is used to describe and analyze machine intelligence in human terms, in our case, however, we argue that true intelligence and agency arises from and within the collaboration between humans and the machine. Hence, it sensitizes us to possible shortcomings of competencies and capabilities arising in co-performance, where consumers act as sensors that need guidance and support by the agent.

### 7.6.1 SENSE: Interact with Food (Waste)

According to Bertran et al. [3] the increased use of automation and sensors leads to an increased agency of the technology rather than encouraging human-food interaction and even might compromise this interaction. Here, our design provides an alternative that encourages more interaction with the material at the border between food and waste. And although we have no insight on an actual food waste reduction, our evaluation shows how the design is perceived to increase the value of food and to encourage conscious embodied interaction with food, which directly addresses current practice theoretical findings [134]. In particular, more experienced consumers agreed on the importance of first-hand experience and the empowerment of the own senses. Hence, a useful and enabling design does not necessarily need more or the newest sensors (e.g., for proof edibility), but leaves room for conscious and independent action. Here our research operationalizes the call of Hassenzahl et al. [133] for more conscious interaction to enhance the experience of and engagement in the practice.

From a co-performance perspective, an agent without sensing capabilities relies on and engages human sense-making. Therefore, the interaction itself reconnects humans and food, which bears broader implications for Human-Food Interaction in the sense to use the agency and limitations of the technology to encourage more agency on the human side. Regarding this, the evaluation highlights the importance of not being patronized by the agent and emphasizes consumers being in control of decisions and sense-making. Complementing visuals or sensors that were discussed to increase reassurance and minimize the risk of a wrong decision, could even impede the sensory training and increase technology dependency. Here, a field of tension between human reliance on technology, bodily reactions, and safety (or efficiency in other contexts) emerges.

In conclusion, the design should encourage the human to use and trust their own senses to build the embodied knowledge they need. For future designs, voice agents should be considered to expand knowledge beyond food waste and motivate the human to appreciate and engage in food interaction. This could be done by incorporating additional information like regionality or seasonality serving the perception of food value [134].

### 7.6.2 THINK: From Machine Knowledge to Human Thinking

From a thinking perspective, it is acknowledged that consumers quickly get confused when trying to rationalize their bodily reactions to the material, which results in the use of institutionalized knowledge [134], in our case the reliance on shelf life and the disposal of food. Regarding this, the evaluation of our prototype shows how Fischer Fritz addresses this problem by providing the means of a step-by-step approach and reassuring the human in his doing. In this sense, the agent takes over part of the thinking, while leaving room for 'sense' on the human side. This distribution of tasks was perceived as increasing confidence in decision making as long as the agent is a trustful entity.

This separation of human sense-making and machine thinking, however, requires a common language. Regarding this, the pre-study already sensitized us for the language used in the specific task of assessing fish that relies on metaphors and figurative language. Our evaluation revealed how the challenge is to balance short commands and carefully verbalized instructions to move the co-performance further without tiring the patience of or confusing users. This shows how mutual reliance and common language in a task allows for collaboration beyond simple tasks [289]. A further aspect of collaboration in thinking is the perception of the agent and his capabilities. Although the agent was compared to humans regarding senses, it was not expected to act or think human-like. Moreover, it fulfilled its purpose by being informative and providing traceable explanations and guidance. In this respect, the machine does not have to mimic human behavior but can complement the human on its own terms [197, 53, 177].

As the participants noted, the agent is ultimately a learning tool which, after the temporary takeover of thinking, needs to provide the means to teach the consumer and finally leave the consumer with its own thinking about the bodily reactions. Active support for reflection, and demonstration of the practices contribute significantly to the transformation from institutionalized knowledge to embodied knowledge as our participants reflected on the prototyping approach. This is in line with the claims of purposive learning and active participation in the practice [109, 354]. And although the machine might take over some thinking, learning always relies on the promotion of self-reflection and the negotiation of (embodied) knowledge that depends on successful human decision-making leading to the experience of self-competence and autonomy.

To further leverage the role of a coach, the voice agent has to ask more open questions, allow for mistakes, and more exploration. Concerning the dialog, that means to allow for intelligent fallbacks that do not feel like dead-end conversations but are enlightening and encouraging [197, 53].

### 7.6.3 ACT: Side-by-Side with an Agent

Thus far, voice assistants do not succeed to engage humans in directed co-performance or conversations [268, 53]. In our approach, the agent and the human have to collaborate and use their unique capabilities to accomplish their goals in practice [109]. The human naturally embodies the use of senses but needs the agent to guide the procedure and classify sensory interpretations. Hence, they complement each other in their distributed capabilities. Usually in human-machine interaction, users act through the machine by direct commands [289]. In our case, both are acting upon the real world through talking and listening and working side-by-side. Interestingly, it is even the agent who leads the interaction of the human with the food. The agent is responsible to communicate the information comprehensibly and adjusted to the humans' capabilities. Yet, the human can decide any time to end the interaction or to just not trust the advice. In comparison to full automation, the human is actively involved in the decision-making process and can control it in reasonable limits. By assigning power to the voice assistant through knowledge and the ability to communicate in human terms, it acts as an equal collaboration partner next to the human [177]. Kuijer et al. [177] claim to not to use human-likeness as an indicator to assess machines. In our evaluation, we could observe that the consumers were not doing that either. Instead, Fischer Fritz met their technological expectations and was judged by its technological capabilities. Future design research should therefore focus on how to adapt human features, like, e.g., showing empathy by using a specific set of words and sounds and transform it into technological terms.

As stressed by Gherardi and Nicolini [108] knowledge means to have the 'competence-to-act' which goes along with engaging in action [133]. To develop embodied knowledge, consumers have to act on their received knowledge, gain experience, and memorize the differences in sensory impressions. Thereby, the voice agent acts as a communicator and offers the human opportunity to link distinct actions with applied knowledge. Thus far, domestic co-performance is often discussed in terms of efficient automation and the elimination of human decision-making [210, 117, 5]. Instead, we have to analyze the gains and losses long-term, when decision-making is completely handed over to an agent. Along with our case study, we could view different levels of consequences when we lack the ability of food quality control. By experiencing the competence to act, similar to the mastership of a former apprenticeship, with every interaction, the human might appropriate the capabilities of the agent and transform deliberate actions into practice. Thereby, our design is not limited to the scenario to prevent food waste but is appropriate to enhance any agency in craftmanship with materials at hand [327]. Consequently, the human is empowered and enabled to act alone at some point, but still has the reassurance to ask for support in any case of uncertainty. We did not follow an approach designated to educational goals, but rather leverage the sense of urgency, as the user rarely decides on planning to tackle the problem of food waste. However, in this respect, the role of the agent can be adjusted, either to support even quicker decisions or checks, or to anchor knowledge and even more learning by additional information.

Finally, both the agent and human, complementing each other by their capabilities, need to engage in collaboration to act upon the world and accomplish their goals. Similar to thinking, the repetitive offer and mentoring of actions contribute to humans to acquire the competence to act on their own and establish new practices.

## 7.7 Limitations

Our study encounters several limitations. Neither did we conduct a formative usability study nor a study in the wild to investigate long-term behavior change, effectiveness of decision-making support or to adjust further critical speech related form factors to ensure smooth interaction by a majority of users. Our aim was to explore the design space by Research through Design with a focus on leveraging the opportunities that come with voice interaction and showcase the design of interactive agents to support domestic practices. Future work needs to evaluate the long-term effects of interaction and appropriation regarding the impact on food waste prevention. Although, we cannot elaborate on the possible effectiveness on footprint reduction of this intervention nor claim that this will impact sustainability on a large scale, we followed the call by Hebrok et al. [134] for more situated consumer decision support along the food lifecycle and offered an alternative approach to persuasive technology design. Furthermore, the lack of cultural comparison is clearly a limitation of our study being grounded in western consumption patterns. Future design studies should address and include culturally related constructs of notions of edibility and freshness.

## 7.8 Conclusion

The present case study proposes the design of a voice assistant which supports the negotiation and transformation of institutionalized knowledge to embodied knowledge to prevent food waste. Our prototype *Fischer Fritz* offers humans a domestic co-performance to decrease personal insecurity and gain the competence to act. Empowering human sense-making and decision-making leads to engaging experience and action without compromising the food relationship. Consequently, this work contributes with its detailed design process to design knowledge as well as to considerations on co-performative sensing, thinking and acting between conversational agents and humans. Future alternative case studies might strengthen the understanding of design practices of interactive agents and learning environments.

# 8 Designing an Interaction Concept for Assisted Cooking in Smart Kitchens: Focus on Human Agency, Proactivity, and Multimodality

## Abstract

Connected homes and smart assistants shape the future practices of humans, but they do not yet perfectly fit their needs and processes. Our research explores how smart assistants can effectively support users during cooking. First, we completed an observational study with ten participants to understand their needs for competence and autonomy in relation to their individual cooking. Following the empirical results, we prototyped a multimodal assistant that interactively provides stepwise guidance for a multi-part recipe. We evaluated the prototype in a Wizard-of-Oz approach with ten participants. The classification according to cooking competence and need for autonomy turned out to be an efficient way to understand the different user perspectives on the prototype. We could observe under which conditions users prefer graphical or voice interaction and how proactivity of the assistant affects human agency and derived general insights for the design and co-performance of smart assistants in other domains.

## 8.1 Introduction

Amid technological progress and predictions of unprecedented growth (cf. [266]) in the smart kitchen sector, current smart home solutions require further development to provide added value to their users. Their current set of functionalities is limited to remote control of appliances and status monitoring via smartphone apps, or improving home security [310, 121, 148]. Consequently, prior research focuses on the advancement of home automation in the context of intelligent assistance, including work on multimodality (cf. [186]), conversational assistance (cf. [113, 60], [207]), and the role of proactivity [268]. According to the data, a reasonable level of proactivity contributes to the user perception of a trustworthy and useful assistant [172]. As a result, designers must examine the implications for interaction design to determine ways to balance proactive systems and human agency. As proactivity increases, a single modality may be insufficient to communicate the system's behavior and the reasoning behind its recommendations and actions. With only a few studies focusing on the optimal combination of multiple simultaneous interaction modalities [324], we lack an understanding of how multimodality might contribute to user-centered proactivity which allows smart home assistants to adapt and act according to the users' information needs [172, 268, 90].

However, other researchers call for alternative visions of smart homes that consider less technophile user needs and emphasize human agency, which positively effects their self-efficacy, autonomy, and competence in their everyday practices [310, 291, 5, 75]. Bertran et al.'s systemic mapping of Human-Food Interaction literature [3] shows that research tends to focus on technological automation. While technology might support and teach humans to perform food practices autonomously, it also risks negatively affecting human food skills and competences in the long term. Some work like speculative design of Human-Food Interactions (HFI) [78] emphasizes human-centered design spaces [327, 14, 55] to explore potential empowerment of human food practices through technology [3], and providing engaging experiences with technology [268]. Future smart kitchens and food environments will require a strong understanding of how human agency, as well as their need for competence and autonomy, influences the design of smart assistance. We define a smart assistant as a system that provides smart assistance to its users by orchestrating smart services and appliances, and facilitating user-centered interaction. This paper explores the following research questions focusing on assisted cooking:

- RQ1: How do users' needs for competence and autonomy influence the design of smart assistance?
- RQ2: How does users' agency influence their assistance needs?
- RQ3: How might multi-modality contribute to user-centered proactivity of smart assistants?

Against this background, we followed a user-centered design approach. We empirically investigated the cooking practices and corresponding resources necessary to prepare a sophisticated recipe. Our observational study, which included pre- and post-interviews with ten participants, led to preliminary design implications that informed our mid-fidelity prototype for proactive smart kitchen assistance. Our respective conceptual design integrates and orchestrates different smart kitchen appliances. It combines graphical and voice user interfaces as interaction modalities to adapt to the user's context. Also, we provided and extended interaction modalities that can meet users' divergent and individual needs with respect to competence and autonomy. For evaluation, we conducted a Wizard-of-Oz study with ten participants to explore the combination of input and output modalities: speech, touch and visuals. Further, we investigated the balance between varying proactivity levels of the assistant and the human need for autonomy and competence. In line with these particular questions, we aimed to understand to what extent users perceive assistance as adaptive, supportive, and pleasant.

After analyzing our participants' cooking competence and need for autonomy, we derived four groups: beginner, accurate cook, creative expert, and creative spontaneous. This classification served as a valuable design rationale for assessing the continuum of the assistant's proactive behavior. The effects on human agency require future kitchen assistants to take the user's level of competence seriously and recognize potential errors early. We could also identify the required complementary interaction modalities, such as speech, touch or visuals, that match user preferences and allow seamless adaptation of communication and information styles. The evaluation reveals the participants' perceptions of control, transparency &

trust, and decision autonomy in interaction with our cooking assistant. With further research, our findings and design implications can be generalized to other domains and used by designers and developers alike. Our work contributes to the development of future autonomous, proactive and multimodal assistance that prioritizes human needs over automation at all costs.

## 8.2 Related Work

### 8.2.1 Smart Everything

Today, smart homes consist of various sensors, actors, and networks (IoT) that promise to automate processes in terms of security, efficiency in routines, comfort, entertainment and leisure [322, 48, 310, 47, 29]. Intelligent Personal Assistants (IPAs) have conquered the home as ubiquitous user interfaces, promising even more personalized and helpful home services [60]. In many cases, off-the-shelf products treat the home primarily as a "technological space" [147] to increase the productivity and multi-tasking of the inhabitants [310, 147]. Meanwhile, not all of them recognize and embrace embedded systems as gaining comfort and smartness [211, 333]. Several studies [211, 75, 90, 5, 268] suggest that people are partly afraid of becoming lazy and passive versions of themselves, missing out on activities they find meaningful and enjoyable.

Particularly smart kitchens are a timely and relevant example of competing autonomy in decision-making, performance, and control of practices [3]. Current IoT systems focus on solving very specific issues or digitally enhancing appliances like content-aware fridges with (semi-)automated replenishment [121]; more integrated solutions include smart home apps provided by home appliance companies, such as Bosch Home Connect [138] or Samsung SmartThings Cooking [277]. These apps offer recipe guidance and meal planning and can wirelessly transfer settings to devices. However, customers only receive a limited amount of assistance, and they still need to take manual actions, e.g., clicking on additional buttons to transfer settings to appliances.

As a side-effect, highly-automated IoT applications may risk diminishing human agency. However, researchers [310, 291, 5, 75] call for alternative visions for IoT applications that consider daily practices and favor human agency over automation. This paper explores how human agency affects users' need for assistance and its design.

### 8.2.2 Interactive and Guided Cooking

Beyond the use of tools and appliances to master the desired recipe, cooking as a practice is based on the creative negotiation of personal preferences and specified instructions [14], often grounded in embodied knowledge [212, 108, 14]. Embodied knowledge refers to prior subjective experience, e.g., cutting techniques with knives or human senses to investigate the freshness of food [14], and intuitive decision-making that is challenging to express and teach through instructions alone [108, 91]. On the contrary, we have also institutionalized knowledge or theoretical rules and approaches that are easier to express explicitly but may not encompass the entire practice. Further, cooking is a part of various human food practices that are integrated, dispersed, and interconnected [184, 343] and, thus, need particular consideration of the users' materials, competences and meaning involved.

Guided cooking has entered the consumer market with the appearance of food processors with cooking functions, such as the Thermomix [165] or Bosch's Cookit [36]. Users can be guided through recipes with detailed step-by-step instructions. This includes automatically setting the correct cooking functions at the respective cooking steps, such as temperature, cooking time, and mixing or stirring level [165], but does not explicitly promote competence building. While these food processors already include guided and intelligent functionality, they are standalone solutions. Further, they neither do address how users can successfully use IoT in the kitchen, nor do they incorporate other connected kitchen items such as ovens or smart assistants. This may impinge additionally on users' frequently poor skills and competences required to cook with traditional tools. Especially IPAs like Amazon's Alexa or Google Assistant are destined to vocally support such practices [89]. However, they are mostly installed as seamless information hubs to control and monitor connected devices in the home [75, 147].

Approaches in academic research attempting to provide holistic assistance concepts are, for instance, MimiCook [278] and KogniChef [229]. Both employ cameras, microphones, and scale sensors to track user activity and collect contextual data for recommendations. Moreover, KogniChef controls connected kitchen appliances autonomously. Nonetheless, a core design principle of KogniChef is to keep users in control and to adapt to them as much as possible. Also, the work of Kato et al. [157] is an early example of interactive cooking to empower users' skills development. [157] investigated the use of visual information for mediating a subjective experience to enhance the comprehension of the effects of physical and chemical processes on food. MimiCook [278] guides the user stepwise through a recipe using activity recognition in an instrumented kitchen. Additionally, a projector displays the instructions via augmented reality. However, this system is concerned with activity and context recognition and has not yet explored user-centered interaction design.

Some HFI researchers [79, 14, 3, 55] emphasize more human-centered design spaces. Dolejšová & Wilde et al. [79] reflected on the anticipation of food tech issues concluding that "technologies can reduce socio-culturally and sensorially rich food experiences into utilitarian, standardized tasks performed by algorithms." Their work emphasized that designers should support the "full organoleptic experience of food" meaning to promote creative and experimental food practices over prioritizing automation-driven convenience. Although recipes are replicable, the performance of cooking is highly intrinsically contextual. Currently, less attention is given to technology that enables humans to develop their skills and competences while avoiding or reducing their technological dependence. Agency in that sense refers to humans' ability and capability to perform food practices by themselves instead of shifting every action and control to technology [3].

Users' feelings of autonomy and competence are an integral part of "'pleasurable experiences' with technology" [131]. Thereby, autonomy refers to being "the cause of your own actions" [131] and competence to being "capable and effective in your actions" [131]. Both terms are closely related to proactive behavior and the control of actions. Consequently, we investigate how the universal psychological needs for competence and autonomy [131, 128] influence the design of smart assistants.

### 8.2.3 Multimodality and Proactivity

As previously outlined in Section 8.2.1, AI-based services tend to advance and leverage autonomous systems in homes. We investigate the continuum of a human-centered system proactivity and its design of interactive components.

Proactivity refers to the autonomous behavior of a system that predicts an upcoming event and initiates a response before the situation arrives [232]. Kraus et al. [173] developed a taxonomy for the proactivity level based on a literature review and a complementary user experience study with a proactive prototype, classifying four levels: "none", "notification", "suggestion" and "intervention". Additionally, Cila et al. [57] propose to assign an agency to IoT artifacts, thus treating IoT as agents that share an interdependence with humans and shape a network of relationships. The authors emphasize that an unresolved distribution of control might "cause a growing tension between human and product agency" and impede a pleasant experience. For the development of proactive systems, the social impact of the level of proactivity and its consequences on the co-performance of practices between systems and humans must be considered [57, 177]. In fact, current research challenges the concept of passive IoT in domestic life to conceive the meaning for the relationship between IoT agents and humans and to formalize implications for design [75, 57, 268, 182, 90]. These approaches focus particularly on the role and capabilities of smart agents and explore the multimodal interfaces for human-computer interaction without compromising human agency.

The potential adaptability and proactivity of IoT systems go beyond their ability to gather information autonomously. On closer inspection, multimodality offer opportunities to enhance the user's learning process of cooking and handling cooking appliances. Earlier work such as "cooking navi" [124] investigated multimodality, which combines cooking video segments, a foot switch as a mouse replacement, and a waterproof touch pen, demonstrating the effectiveness of the provided multimedia assistance. However, conversational assistants had not been considered or explored at that stage. Furthermore, no results were reported on how participant characteristics influenced the use of the assistant, although participants differed in their cooking experiences. Wechsung and Naumann [345] show that offering multimodal interaction can increase the perceived user experience of an application, although a single modality would be more time-efficient in accomplishing the task. According to Schaffer and Reithinger

[279], the use of multimodal interaction is currently uncommon. However, this is due to the high complexity of developing and a lack of experts, not a lack of purpose. They explain that combining the strengths of the individual modalities might increase the system's efficiency, robustness, accessibility, and intuitiveness. Nevertheless, user effort to interact with multimodal applications must be minimized [279]. Similarly, Lazaro et al. [186] emphasize to adapt the modality of intelligent systems to tasks and user characteristics. However, their literature review reveals a lack of empirical studies focusing on the interaction between the various modalities and human factors.

By empirically investigating the context of cooking guidance, we intend to gain in-depth knowledge on designing proactive, empowering, and smart assistants while preserving a balance with human agency. Therefore, we will focus on proactivity and multimodality as key design factors.

## 8.3 Observational Pre-Study: Methodology

As a first step in the user-centered design process, the observational study aimed for a thorough understanding of the user context of assisted cooking. Therefore, the pre-study was conducted as a qualitative participant observation (cf. [301]) during a cooking task.

### 8.3.1 Method

**8.3.1.1 Study Design and Procedure** The observational pre-study consisted of three parts: a semi-structured pre-interview, an assisted cooking observation and a semi-structured post-interview and was conducted in a smart kitchen lab, the ”Smart Life Lab” at Bosch's research campus in Renningen, Germany, equipped with a set of high-end connected smart kitchen appliances, notably including a Cookit [36], connected oven and stove, and recording facilities for audio and video. Participants were provided with all the necessary ingredients. Prior to cooking, we briefly introduced the kitchen appliances to familiarize them with the lab kitchen, e.g. setting the oven and selecting a recipe for the Cookit. The appliances offered the standard range of off-the shelf functionality, e.g. oven programs, step-by-step recipes of the Cookit. However, they were not connected to an app.

Our pre-interview before the cooking task focused on understanding the participants' previous cooking experiences, appliances used at home, cooking challenges, approach to information gathering, and handling of recipes. For the assisted cooking observation, we deliberately chose a challenging recipe to uncover information needs and to investigate tool handling for potential assistance needs: “Asparagus in Pancakes gratinated with Sauce Hollandaise”. This recipe requires intermediate skills and high attention, e.g., several interdependent steps, the peeling of asparagus, making a hollandaise sauce from raw ingredients, and several smart kitchen appliances. We did not hand out a recipe or provide any details about the process or ingredients prior to the actual observation. Instead, the experimenter acted as a cooking

**Table 6**
Characteristics of observation study participants, sorted by cooking competence (1 = low, 7 = high)

| Participant | Gender | Age | Cooking Competence | Need for Autonomy |
|---|---|---|---|---|
| OP02 | f | 27 | 1.29 | 2.60 |
| OP09 | f | 20 | 2.86 | 2.20 |
| OP06 | f | 22 | 3.57 | 3.40 |
| OP04 | m | 54 | 3.57 | 2.00 |
| OP03 | f | 26 | 3.71 | 5.20 |
| OP10 | m | 29 | 5.29 | 3.00 |
| OP07 | f | 57 | 5.57 | 4.60 |
| OP08 | m | 63 | 5.57 | 4.20 |
| OP05 | f | 57 | 6.14 | 4.40 |
| OP01 | f | 32 | 6.71 | 5.40 |

assistant [82] with the knowledge of an ambitious home cook. The experimenter invited the participants to ask her any questions regarding the recipe, cooking appliances, and food preparation, to uncover their personal information and guidance needs during the study. We instructed the participants to make their own decisions about how to cook and what materials and devices to use, but implied to ask for advice when needed. Besides, we asked the participants to "Think Aloud" [85] throughout the cooking process. All experiments were conducted in German. Afterwards, we conducted a post-interview that first addressed the participants' reflections on the cooking process, including challenges and positive aspects, and their attitude towards smart technology. Secondly, we discussed their use of smart technology, desires for a kitchen assistant, preferred interaction modalities, and expectations for assistant proactivity.

**8.3.1.2 Recruitment** To recruit a sample, we used a questionnaire focusing on cooking competence, cooking preferences and technology affinity. Along with demographic information, the questionnaire inquires about the participant's daily cooking routine, e.g. frequency of preparing a main dish from basic ingredients, enjoyment of cooking and whether the participant cooks mostly alone or with others. The psychological need for competence and autonomy is assessed by an adaptation of items from Sheldon et al. [293] to the cooking domain. The questionnaire is extended with questions from [13] developed to assess the participant's cooking competence. We invited ten out of 30 respondents with different levels of "cooking competence" and "need for autonomy" to participate in the observational study. We aimed for a heterogeneous sample (cf. Table 6) and thus selected five participants with a score below 4 (the midpoint on the 1–7 scale) and five above for each of both criteria. Apart from this constraint, the selection of participants was pragmatic. The average age was 39, seven participants were female, three were male, and all from a different educational and professional background, none of whom were professional chefs.

**8.3.1.3 Data Collection & Thematic Analysis** Finally, we examined the verbatim transcripts of the pre- and post-interviews. Additionally, we transcribed relevant paragraphs of the videos to obtain direct quotes from the participants and note our observations of their approaches to cooking and handling tools and appliances. Next, we followed thematic analysis as adapted from [317] to code and cluster our observations and the participants' quotes. At first, we used pre-defined codes that were related to our main questions from the interviews and the research questions to deductively code paragraphs of the transcripts, as follows: needs for autonomy and competence, interaction modalities, cooking actions of the participants incl. whether they diverged from or followed the recipe steps, as well as information needs. Afterwards, codes emerged inductively within the paragraphs of the main codes. For example, to further understand the "recipe behavior" of the participants as a main code, subcodes like "follows recipe instructions" or "changes recipe" were used. By reviewing each code separately, we could use the data to classify our cooking types and analyze which support and information needs might be useful to them. Additionally, we analyzed the challenges experienced by the participants for future assistance technology design. The same process applies for the video analysis with a particular regard to observing the step-by-step process of cooking. For example, one of the main codes was "information needs" and it was further divided into subcodes like "recipe overview", "recipe sequence" or "ingredients". All quotes of the participants were translated from German to English.

## 8.4 Observational Pre-Study Findings: Cooking Types & Interactive Information Support

### 8.4.1 Cooking Type User Classification

The results of the pre-study confirmed our consideration of users' cooking competence and need for autonomy as relevant variables in determining the requirements of cooking assistance imposed by different user types. Accordingly, we clustered our participants along these two dimensions in Figure 8, divided by a threshold score of 4 (the middle value on the scale 1–7), resulting in four types: "the beginner", "the accurate cook", "the creative expert", and "the creative spontaneous". We made two exceptions in the assignment of OP04 and OP08 since we could identify disparities between their results on the recruitment questionnaire and the observational study. In the recruitment questionnaire, OP04 scored slightly below the cooking competence threshold but was deemed competent according to the pre-interview. OP08 indicated a low need for autonomy in the pre-interview, despite scoring slightly above-threshold in the recruitment questionnaire.

Notably, we found some representatives in our sample for all four categories in Figure 8, although there is only one for "creative spontaneous". In fact, it is known from other domains of expertise, e.g., language learning, that greater competence correlates with greater autonomy [70], so the correlation between cooking competence and the need for autonomy is not surprising. This would lead us to conclude that there are mainly two types of users:

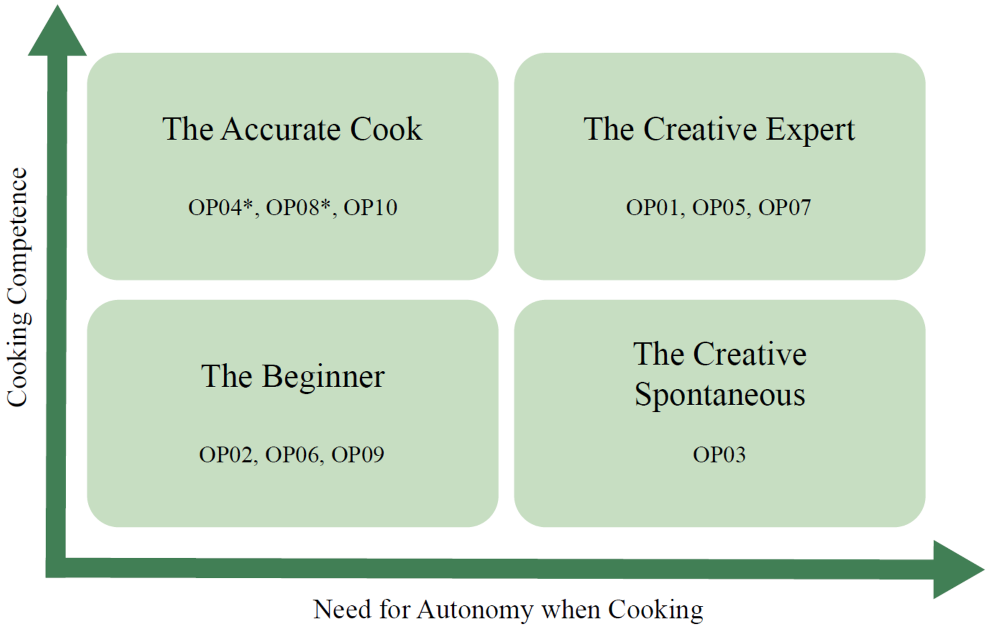


**Figure 8**
Cooking types matrix, * marks reclustered participants

Those with low competence and a low need for autonomy, and those with high competence and a high need for autonomy. However, as we know from language learning, this correlation cannot be expected to be perfect, and here we found participants who were also representative of the other quadrants. This allows us to study at a qualitative level what makes these users different from the two main groups of "beginners" and "creative experts". It should be noted, however, that the small number of participants does not allow us to determine the exact number of these groups among all potential users of cooking assistants. In the following, we describe the characteristics of the four different cooking types in more detail.

**8.4.1.1 The Beginner** The first cooking type (represented by OP02, OP06 and OP09) is characterized by a low cooking competence and a low need for autonomy in cooking. This type includes people who cook infrequently and therefore do not consider themselves competent in preparing a meal: "When I cook new things, then actually always with a recipe, because I simply can't do it any other way" (OP06). Due to their low cooking experience, they are likely to follow the recipe very strictly to avoid mistakes : "I actually always use a recipe." (OP09) The study results support this assumption as all of the three participants followed exactly the recipe order and the advice by the experimenter in the role of the assistant: "I'm overwhelmed when it comes to this kind of thing [...] I just followed your instructions exactly" (OP02).

Additionally, the participants demonstrated a lack of cooking motivation. They refer to insecurities while cooking which might prevent them from gaining cooking competence as they might not gain any positive cooking experiences. Consequently, an appropriate assistant should aim to reduce users' insecurities and increase their motivation to cook, e.g., by providing detailed instructions including tips on unknown tasks. This is supported by the appreciation of instructions being step-by-step (OP09) and "direct" (OP02, referring to the guided cooking provided by the Cookit).

**8.4.1.2 The Accurate Cook** The second cooking type (represented by OP04, OP08 and OP10) is very accurate in following recipes and exhibits a low need for autonomy. However, this group is characterized by a high level of cooking competence. People in this group cook regularly and are therefore experienced in the usual cooking activities. In contrast to *creative experts* (discussed below), they do not feel restricted by precise instructions. OP08, for instance, notes that the Cookit "always says what to do" which he deems "totally practical". In contrast to the *beginners*, the pre-interviews suggest that the *accurate cooks* are more likely to adapt recipes when it becomes necessary, for example, when the corresponding ingredient is not available. This aspect seems to suggest that they can perform most steps independently and without additional information, due to their cooking experience. Hence, a matching assistant type might support them with clear instructions, but without bothering them with too much detailed information.

**8.4.1.3 The Creative Expert** The third cooking type (OP01, OP05 and OP07) is characterized by a high cooking competence and a high need for autonomy. These people have significant cooking experience and show a creative and spontaneous way of cooking, often without any recipe. This group includes the participant with the least amount of questions during the observation, OP05. Unlike the *spontaneous creatives*, the *creative experts* rely on their experience when following or changing a recipe, as OP1 puts it: "So if things don't fit for me as it says in there, then I just make it fit, because the recipes are not always all correct." who goes on to explain an example of an incorrect recipe. Comparing the statements of this group with the group of *accurate cooks*, they feel more constrained by detailed instructions like the ones Cookit gives. Based on these results, we expect that people from the group of *creative experts* need more flexible support. It seems essential not to limit their creativity and to provide assistance that enables them to make their own decisions. For example, OP07 claims that sticking to the recipe is "not that important" and that she "got annoyed" with the Cookit.

**8.4.1.4 The Creative Spontaneous** The last cooking type is characterized by a high need for autonomy but low cooking competence. Creative spontaneous cooks are more likely to modify recipes to fit their mood and add their own twist. They dislike following precise instructions. There is only one participant in the observation study who falls into this category (OP03). Her competence score of 3.71 is also just below the threshold, so the following discussion of the cooking type *creative spontaneous* should be considered with reservation.

However, in line with her characterization as creative spontaneous, OP3 did not ask for instructions before starting to cook, and in the interviews, she recalls situations where assistance would have been beneficial, e.g., when she “forgot something important”. Nevertheless, she indicates feeling “hampered” by the order of the recipe steps and restricted by the detailed instructions of the Cookit. Given the low level of cooking competence, it can be assumed that the cooking experience of creative spontaneous people does not allow them to reliably assess the effects of recipe changes. This would lead to failing to achieve the desired cooking results. In fact, OP03 describes being aware of situations that support this assumption. These characteristics make it difficult to assist such users. Based on these results, an assistant could act with restraint but recognize mistakes in cooking, prevent them, or at least assist to correct them. This requires appropriate sensor technology to detect errors and ensure that only need-based support is offered.

### 8.4.2 Cooking & Information Support

Below we present the various information needs identified throughout the cooking process.

**8.4.2.1 Information Search Behavior** Overall, participants search for classic recipe information on the internet using Google as well as in cookbooks. Mostly, they seem to search for entire recipes that are bookmarked in smartphone apps such as *Kitchen Stories* [167] or *Chefkoch* [50]. If a particular step in a recipe is unclear, or the preparation of a particular food is unknown, they usually tend to search for that information online. In case of remaining uncertainties, the participants also watch short videos of the said cooking method, e.g., on YouTube.

**8.4.2.2 Classical Recipe Information** In the following, we summarize the questions directed to the Cooking Assistant during the cooking process.

**Recipe Overview & Scheduling** When starting to cook a recipe, usually as the first step, participants review the recipe overview to learn about the individual steps and their relation to each other. In addition to questions about ingredients and their quantities, some participants asked for explanations of quantities such as “a pinch”. In most cases, a cooking process has a specific, predefined sequence of steps. The observation has shown that some of the participants have decided on this sequence on their own. Often, they thought about the optimal order of steps considering their expected duration; sometimes they decided spontaneously how to proceed. Some participants stated that the order of the steps does not match their personal preferences, so in practice they often restructure their recipes. Timing, e.g., multitasking and concurrency, is also mentioned as major challenge.

**Food preparation** In particular, the preparation of certain foods, e.g. asparagus, often resulted in participants asking for instructions, e.g. preparation method and cooking time. Occasionally, participants asked about the reasons for applying a particular method, e.g. why eggs need to be separated. Even background information that is not necessary to achieve a desired cooking result was requested and considered relevant.

**Kitchen appliances and utensils** A number of kitchen appliances were needed, and often there is not just one appliance that can be used for a particular cooking task. Therefore, some of the participants asked for help in choosing an appliance, e.g. whether to use a hand mixer or the Cookit. OP04 expressed selection criteria such as low effort. Participants also asked about appropriate settings, e.g., preheating the oven. Moreover, they had questions about the use of appliances that they had never used before, e.g., the Cookit.

**Demand for Preventing and Correcting Errors** One of the most prominent challenges for the participants was the handling of errors that occurred during the cooking process (which our experimental setup allowed us to observe). There are different categories of errors: those that are technically easy to detect, e.g. forgetting to start an appliance, and those that would require highly intelligent sensor technology in the kitchen. Additionally, some errors can still be corrected afterwards, while others have irreversible consequences, e.g., putting a whole egg into a mixture instead of egg whites. Regardless of the type of error, participants want a smart kitchen assistant to help detect errors in time or, if that is not possible, to help manage the consequences properly.

### 8.4.3 Multimodal Interaction

In our analysis we could identify appropriate interaction modalities for assisted cooking in smart kitchens. All participants explained their preferred and acceptable interaction modalities while cooking as gestures, voice, graphical interfaces, and combinations thereof.

**8.4.3.1 Gesture-based Interaction** Exactly half of the participants (OP01, OP03, OP04, OP06 and OP07) considered gestures to be a good option (OP01, OP03, OP04, OP06 and OP07), particularly because of the potential for dirty hands from cooking (OP01 and OP03). Additionally, OP06 requested confirmation feedback to clarify whether the interaction was successful. The other half of the participants indicated that they would probably refrain from using gesture control because of the lack of benefits due to the difficulty of matching gestures with spontaneous movement patterns.

**8.4.3.2 Voice User Interfaces** Our participants discussed voice interaction as a potential input and output modality, concluding it might be convenient for cooking. Similar to the case of gestures, dirty hands during cooking are mentioned as a reason (OP03, OP06, OP07, OP09). Furthermore, OP01 and OP10 stated that they regularly use voice assistants at home. However, OP07 and OP04 were concerned about possible misunderstandings by the voice assistant when engaged in noisy cooking activities. The results for voice output clearly show that most participants (8 of 10) can imagine using voice as an output modality in the kitchen. For example, OP02, thinks that the human-like effect could even motivate to cook. OP01 and OP03 would not use voice output in the kitchen — partly because of the loud noises in the kitchen.

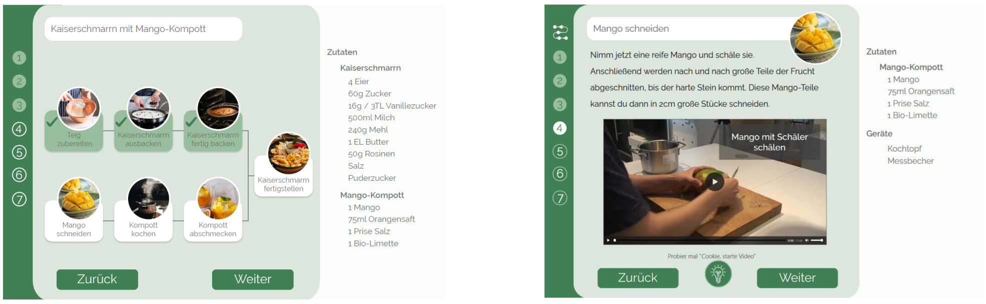
**8.4.3.3 Graphical User Interfaces** Apart from one participant who left an answer blank, the majority of participants (8 out of 9) thought that a GUI combining visual output and touch input was a fitting interaction. OP01, OP04 and OP08 hold this view because they are already experienced in using GUIs. OP03 and OP07 describe that they especially need a visual component to understand the complicated issues of cooking. In contrast, OP07 cannot imagine using a touch screen because of the dirty hands while cooking.

### 8.4.4 Considering Proactivity

In light of users' need for autonomy and their statements, we can derive some implications regarding the proactivity levels of the assistant. The majority of participants (five out of six) who were classified as having a low need for autonomy, think in general it could be helpful to have a proactive assistant for cooking, e.g., as OP08 puts it "if you don't ask the right questions, you won't get an answer". However, they are also concerned by its level of proactivity, e.g., only if "proactivity doesn't happen excessively". OP02 would like to be assisted proactively only with issues that are essential for the success of the recipe, but purely on demand otherwise. OP10 would prefer to get proactive support from an assistant only for a very good reason, e.g., if a crucial ingredient is still missing. By contrast, none of the four participants classified as having a high need for autonomy consider a proactive assistant to be generally beneficial. Whereas two participants consider proactivity to be "annoying", one participant (OP3) would accept proactivity if the assistant is capable of determining whether she is ready to be assisted. OP07 points out that she appreciates proactive support depending on her mood. A precondition for her would be a configuration option to turn it off. In summary, our findings highlight that the provision of proactivity requires an accurate assessment of individual users' needs (taking into account their current intention and situation). When interviewed, some of the participants confirmed the importance of assistance concepts for the kitchen being designed holistically, for supporting the whole cooking process.

**Table 7**
Design implications for the prototype

| |
|---|
| **The Assistant.** The observations established that *step-by-step* guidance is needed. This includes help in choosing kitchen appliances and tools, and their operation. A good *overview* of the recipe is required. For beginners, a sufficient level of *details* and *hints* is crucial. Providing *choice* and *explanations* is seen as valuable for building user competence and trust. However, acting in the role as assistant has revealed the difficulty to intervene in cooking processes when mistakes happen but users won't ask for advice. Therefore, hints should be categorized by priority and severity to the success of meal preparation. |
| **Multimodal Interaction.** Among the various interaction modalities under consideration, the combination of a GUI with touch input and voice interaction was determined to be the most promising approach for interaction and instruction. Thereby, we will differentiate between in- and output and the priority of information. |
| **Flexible Navigation** Users should have the flexibility to navigate back and forth and choose what information they want to engage with, based on their needs and personal cooking process. |
| **User-centered Proactivity.** The identified user types differ in their need for *details* and *hints*. The latter should be provided only when needed, especially for more experienced users. This implies making hints available on demand, allowing users to select/navigate hints, and enabling them to ask questions. However, if the user does not ask for help, then assistance must be offered *proactively* by an assistant (system) in the event that an error is detected, can be prevented, or needs to be corrected. By being the assistant ourselves, we could observe unsolicited advice is quickly perceived as inappropriate and not justified. Yet, to give the opportunity to learn, sometimes patience is more important than preventing errors. |



Kaiserschmarrn mit Mango-Kompott

Zutaten

Kaiserschmarrn

- 4 Eier
- 60g Zucker
- 16g / 3TL Vanillezucker
- 500ml Milch
- 240g Mehl
- 1 EL Butter
- 50g Rosinen
- Salz
- Puderzucker

Mango-Kompott

- 1 Mango
- 75ml Orangensaft
- 1 Prise Salz
- 1 Bio-Limette

Mango schneiden

Nimm jetzt eine reife Mango und schäle sie. Anschließend werden nach und nach große Teile der Frucht abgeschnitten, bis der harte Stein kommt. Diese Mango-Teile kannst du dann in 2cm große Stücke schneiden.

Probier mal "Cookie, starte Video"

Zutaten

Mango-Kompott

- 1 Mango
- 75ml Orangensaft
- 1 Prise Salz
- 1 Bio-Limette

Geräte

- Kochtopf
- Messbecher

**Figure 9**

Overview screen (left) and recipe step description (right). The steps shown on the left translate as: prepare dough, bake Kaiserschmarrn, finish baking Kaiserschmarrn, cut Mango, cook compote, season compote to taste, finish Kaiserschmarrn. The listed ingredients ('Zutaten') are: 4 eggs, 60g sugar, 16g vanilla sugar, 500ml milk, 1 Tsp butter, 50g raisins, salt, icing sugar, 1 mango, 75ml orange juice, 1 pinch of salt, 1 lime. Buttons are marked 'back' ('Zurück') and 'continue' ('Weiter'). The instruction text shown on the right provides a detailed description on how to cut a mango.

## 8.5 Designing an Interaction Concept for Smart Kitchens

Based on the results from the observational study and related work, we derived a holistic interaction concept for assisted cooking. In line with an iterative user-centered approach and our research goals, we developed the prototype to be operated in a Wizard-of-Oz fashion, instead of an early functional full development. With our mid-fidelity prototype assistant "Cookie" we aim to explore and evaluate the multimodal interaction design as well as its user experience. Conceptually, Cookie guides through a recipe and orchestrates all appliances and information needs of the users. The guidance and respective advice, decisions and information are following the structure of the recipe, precisely the beginning and end of the activity of cooking which is closely related to the insights from the video analysis. Next, we decided which functions of the smart appliances to integrate and used the respective control app to execute the autonomous events by our prototype Cookie. In parallel, we designed the role and capabilities of Cookie to process the commands and assistance needs of the users as multimodal and interactive in- and outputs. We used AdobeXD to prototype the GUI implementation and Google Dialogflow for the voice interaction. Cookie's voice output was generated by a synthesized female voice. In summary, our prototype is based on the design implications summarized in Table 7 grounded in the results of our pre-study regarding (1) navigation & assistance, (2) multimodality, (3) flexibility, and (4) automation & proactivity . The following sections provide further detail.

### 8.5.1 Navigation & Assistance

All previously identified information needs were integrated into the design of the assistant, including information search behavior, classical recipe information such as a recipe overview, ingredient list, food preparation methods, and additional preparation advice and tips. Following the searching behavior practices, each step is explained using multiple media types such as text, images, videos and auditory explanations.

A graphical flowchart presents a recipe overview that also serves as a progress indicator. It offers users control at certain decision points, e.g. whether to start the recipe with the preparation of the Kaiserschmarrn or the mango compote. (cf. Fig. 9 (left)). Further, the system illustrates each recipe step with text, videos and hints (cf. Fig. 9 (right)).

To resolve the target conflict between necessary information and bothering intervention by the assistant (cf. Section 8.4.4), the hints were classified into three prioritization levels according to their importance. Priority one includes information important to taste and consistency. This information must be provided proactively to the user. Therefore, it is displayed directly in the step description or communicated via the voice interface. Priority two involves hints for easy and practical food preparation. This information is not critical for a perfect cooking result, but it helps to do so in an effective way. These hints are explained through video or voice interaction. The last priority entails information that is good to know, but not essential for the recipe preparation, like more information about organic limes. These hints are only visible to users when they ask for hints or click the “hint" button at the bottom of the screen. This classification and its implementation follow the context-sensitive guidance principle proposed by Neumann et al. [229].

### 8.5.2 Automation & Proactivity

Literature research [157] suggests that users will perceive an intelligent assistant for complex tasks, like cooking, as more trustworthy and competent if it acts proactively. This is in line with our results, which suggest a holistic concept that integrates the different kitchen appliances. Therefore, the interaction concept is designed as a highly proactive approach. Further, the proactive design is closely tied to the structure of a recipe and its components, such as available and required appliances, trigger actions, multimodality of interactions, etc. It can be modified to each of the recipes in the database. Proactivity is defined as the system monitoring the status of the kitchen appliances and the cooking process and being able to control the appliances in a predictive and demand-driven manner for the user, e.g., preheating the oven. The proactive events have different triggers to detect the user’s intention.

This can either be a user input, such as finishing a previous step, or a status change of appliance data, such as an expired timer. Each event triggers specific actions to advance the cooking process. The triggered action can consist of either giving a recommendation to the user or controlling or adjusting appliance settings. An overview of the relevant events is shown in Table 8. Note that the first event E1 is considered proactive with the oven activated ahead of time without a user request, even though its trigger is a user action. The events are executed through a Wizard-of-Oz action via Dialogflow for voice interactions, GUI remote control, and Home Connect actions for the appliance settings.

When designing proactive events, we followed research-based design principles (e.g. [365]).

**Table 8**
Overview on prototype Cookie's proactive events in the user study

| Trigger | Action |
|---|---|
| E1. User selects first step for Kaiserschmarrn (GUI or VUI) | Oven: is preheated to 220° top and bottom heat<br>VUI: informs user about event |
| E2. User decides to use Cookit to prepare Kaiserschmarrn dough (GUI or VUI) | Cookit: is switched on and the recipe for Kaiserschmarrn is started<br>VUI: Informs user about starting the Cookit |
| E3. Step 18 in Cookit recipe is reached (Cookit GUI input), which means the Kaiserschmarrn dough is finished | GUI: screen switches to baking the Kaiserschmarrn<br>VUI & GUI: user is asked to put the pan with some butter on the stove |
| E4. User puts pan on the stove, Wizard recognizes pan on stove | Stove: perfect fry roast sensor level 3<br>VUI: Informs user about heating the pan |
| E5. Perfect fry roast sensor recognizes that stove temperature is reached | VUI: user is asked to fill in the dough into the pan; stove: timer for 4 minutes starts |
| E6. Stove timer is expired | VUI: user is asked to put the pan in the oven |
| E7. Oven is opened and closed again | Oven: timer is set to 15 minutes |
| E8. Oven timer is expired | VUI: user is informed that timer is finished with a hint that pan is very hot |
| E9. Ingredient is missing, which is recognized by wizard | VUI: user is informed that the ingredient is missing |
| E10. Mango is added to the pot on stove, which is recognized by wizard | Stove settings: stove level 6, timer for 7 minutes<br>VUI: Informs user about event |
| E11. Stove timer is expired | VUI informs user about finished timer and asks the user to taste it |

In the case of trigger conditions, the user is directly informed about the event by speech and if necessary by a visual representation on the GUI. This follows the principle of explaining why a proactive action is happening to encourage trustworthy prototype perception. It is important to consider the user's situation and anticipate user needs that equally contribute to error prevention and user protection.

### 8.5.3 Multimodality

The interaction concept is based on multiple interaction modalities entailing input and output preferences. The prototype uses voice interaction combined with a visual component, such as a GUI for touch interaction and visual feedback. The information required for cooking can become complex and overwhelming for the user, so visuals should always be available on the GUI to illustrate instructions. This will also ease memorizing the necessary information. However, we visualized and organized the information according to the previously identified needs (cf. Section 8.4.3) into different types of screens: overview, decision screen with recommendation, step descriptions, and hints. Our GUI is complemented by the Cookit GUI, as users might decide to prepare the dough with Cookit. Furthermore, as suggested by the literature [197] and the preliminary study (cf. Section 8.4.3), a voice user interface (VUI) is integrated into the design. Particularly to motivate less experienced users, voice interaction is designed to be affirmative and rewarding, such as confirming the user "this is a good choice."

The interaction concept ensures that modalities complement rather than replace each other [156]. On the one hand, the visual interaction channel can compensate for the limitations of voice output. On the other hand, users are enabled to use voice input when their hands are occupied, and to switch to touch input when the cooking process becomes too noisy for voice input.

### 8.5.4 Flexibility

Considering users' need for autonomy (cf. Section 8.4.1), the system offers the user a choice of alternative recipe paths, and if there are several equally suitable cooking methods or devices, it recommends e.g. the easier, faster, or more successful one. For people with little cooking experience, this recommendation will be conveyed through both GUI and VUI. On the GUI, alternative options, are presented side by side, e.g., using the Cookit vs. hand mixer. The recommended option is marked with a chef's hat icon as a visual cue and complemented by a brief description of the benefits. Both options are accessible by VUI and GUI. For even more flexibility, users can jump between recipe steps at any time.

## 8.6 Prototype Evaluation: Methodology

In line with a user-centered design process, we used Wizard-Of-Oz [247, 82, 175] as a common method in HCI to evaluate intelligent systems and assistants. This qualitative approach aims to uncover user expectations regarding assistance needs and explore the user experience of our holistically integrated kitchen assistant enabled by a multimodal interface. After outlining our methodological approach, we present the main themes of our user evaluation.

### 8.6.1 Method

**8.6.1.1 Study Design and Procedure** The evaluation had three main parts: a semi-structured pre-interview (see Section 8.3), a Wizard-of-Oz test and a semi-structured post-interview. During the Wizard-of-Oz test the participants had to prepare "Kaiserschmarrn with mango compote" that is one dish but divided in two components with each of them requiring fresh preparation. We decided for this recipe based on several reasons to study the participants' reactions and interactions: It is a moderate cooking challenge, e.g., peeling and slicing a mango, while fitting into a 90-minute timeframe, and having multiple cooking steps. Further, by not restricting and pre-programming the order of execution, it allows for the assistant's proactivity and automation/orchestration of appliances, such as the stove, a hand mixer or the Cookit, and the oven. Its flexible order also enables the participants to make their own decisions or to follow the assistant's advice. Neither of both decisions will affect the final result, as it is irrelevant which component is prepared first. We conducted the study in our smart kitchen lab (cf. Fig. 10), asking the participants to "Think Aloud" during the cooking session. The primary task was to follow and complete the full recipe without further specification of goals or tool use. Compared to the observation study, our prototype (see Section 8.5) replaced the experimenter. The wizard acts as the assistant and controls every

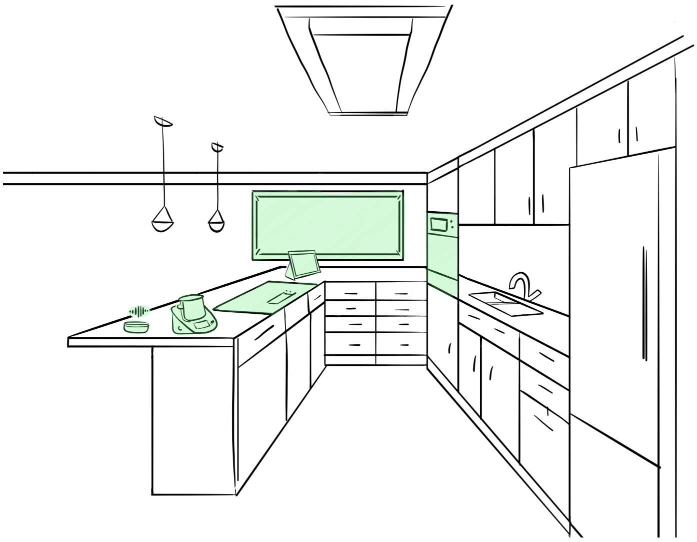


**Figure 10**
Schematic drawing of lab and prototype setup with all smart appliances highlighted.

multimodal component and smart function to mediate the recipe, instructions, and hints by observing and processing the participants' verbal accounts, tool use, and current recipe steps. We have listed the set of proactive events in Table 8 in Section 8.5.2 that are now executed by the wizard. The semi-structured post-interview focused on the participant's experience of the cooking process, assistance, and emerging interactive challenges. We evaluated the proactive interventions of the assistant performed without a user request, e.g., autonomous events triggered by kitchen appliances, the perception of decision moments, and their impact on the sensation of autonomy or restriction. Further, we discussed the different modalities of presenting instructions, e.g., hints based on videos, GUI, or voice. Lastly, we asked for potential areas of improvement and future design ideas.

**8.6.1.2 Recruitment** A sample of ten study participants was recruited using the same questionnaire as described in Section 8.3, covering the topics "cooking competence" and "need for autonomy when cooking". To avoid bias, we also selected "technology affinity" as an additional recruitment criterion and used a subset of questions from the Technology Readiness Index (TRI) 2.0 [241] to measure attitudes towards technology. Therefore, we included two items from each of the TRI areas 'Optimism', 'Innovativeness', 'Discomfort' and 'Insecurity' with a seven-point agreement scale. A total of 39 persons answered the questionnaire, of whom we invited ten (cf. Table 9) to participate in the Wizard-Of-Oz study based on the criteria "cooking competence", "need for autonomy" and "technology affinity". The selected participants (seven females, three males) have different educational and professional backgrounds. These criteria ensured a diverse and heterogeneous sample for the study. All participants received a 20€ voucher, regardless of the outcome of the study. Notably, EP01, EP04, and EP06 already participated in the observational study. However, as no version of the designed interaction concept was included in the pre-study, this is not a limiting factor. The cooking type classification in Table 9 draws on the results of the observational study. Note that two participants with a cooking competence score close to the threshold of 4 were reclassified (marked with * in Table 9). Besides, we had no control over the execution of the self-reported behavior by the participants but carefully recruited to increase the likeliness of participants showing the expected behavior. However, in our analysis we observed only two deviations: EP03 showed a higher level of cooking competence than expected from already knowing the recipe. EP06 is classified as having lower cooking competence resulting from the in-person interview.

**Table 9**

Evaluation participants with scores ordered by cooking type (1 = low, 7 = high; $^{\diamond}$ participants from observation study; * reclustered participants)

| Participant ID | Gender | Age | Cooking Competence | Need for Autonomy when Cooking | Technology Affinity | Cooking Type |
|---|---|---|---|---|---|---|
| EP04$^{\diamond}$ | f | 27 | 1.29 | 2.60 | 2.75 | Beginner |
| EP01$^{\diamond}$ | f | 20 | 2.80 | 2.20 | 5.38 | Beginner |
| EP08 | m | 23 | 3.57 | 3.40 | 5.63 | Beginner |
| EP03 | m | 34 | 3.67* | 3.40 | 5.38 | Accurate Cook* |
| EP10 | f | 54 | 5.43 | 2.00 | 4.63 | Accurate Cook |
| EP02 | f | 33 | 5.00 | 5.00 | 6.00 | Creative Expert |
| EP09 | f | 47 | 5.29 | 3.80 | 3.13 | Creative Expert |
| EP07 | f | 58 | 5.86 | 4.80 | 2.7 5 | Creative Expert |
| EP05 | m | 27 | 6.00 | 5.00 | 4.13 | Creative Expert |
| EP06$^{\diamond}$ | female | 26 | 3.71* | 5.20 | 3.88 | Creative Spontaneous* |

**8.6.1.3 Data Collection and Analysis** The transcripts of the interviews and video recordings were coded inductively using the tool MAXQDA 2020 [203]. As described in Section 8.3.1.3, we used thematic analysis oriented towards [317] and discussed the codes subsequently within the research group. All quotes were translated from German to English. We present our findings introducing the codes as subheadings, as follows.

## 8.7 Evaluation Findings: Multimodal Kitchen Assistant

### 8.7.1 Users' Decision Autonomy

Results from the observational study suggest that the four identified cooking types (*beginner*, *accurate cook*, *creative expert*, and *creative spontaneous*) require different levels of support and autonomy by the assistant. In fact, the three participants classified as beginners appreciated the step-by-step guidance and especially the hints from "Cookie". Faced with a choice, they followed exactly the assistant's recommendations, marked by a chef's hat. EP08 comments: "It's good that it tells me what to start with, because otherwise I wouldn't have decided." Some statements suggest that beginners do not consider being offered alternative paths crucial to their autonomy, e.g. "I'm happy when someone tells me what to do when I'm cooking." (EP04). However, beginners do not follow instructions blindly. These participants also highlighted the importance of background information about recipe steps and decisions to learn something new. With in-depth information backing-up the decision moments, beginners are able to make an informed choice based on the context, e.g., cooking time. Two of the beginners (EP01 and EP04) specifically expressed their dislike of cooking, while EP01 also felt that an assistant could be motivating and make cooking more fun.

By contrast, the assistant for *accurate cooks* (EP03 and EP10) has to find the right balance between proactively supporting the users without bothering them, due to their greater experience in cooking. Unlike the beginner group, these participants often cognitively evaluated the assistant's suggestions instead of blindly following the assistant's advice. Another effect of the high cooking competence of the *accurate cook* group is the way they adjust the recipe on their own during the cooking process. For example, based on his cooking experience, EP03 caramelizes the Kaiserschmarrn in a way that deviates from the recipe. From the standpoint of the creative cooks, assistance seems to be more desirable for new, unfamiliar recipes, since they hardly need help with recipes they know well.

In contrast to the two previous groups, *creative experts* do not always cook according to recipes; they rather use them for guidance and adapt them to their cooking experience. As expected, they showed a higher need for personal autonomy in terms of assistance. For example, EP09 did not follow the assistant's recommendations, preparing the mango compote first and using the hand mixer instead of the Cookit. Still, even creative experts rated the assistant's reminder and task take-over features as desirable, as long as they can still modify them independently. Part of the reason appears to be that they are seeking total control over the cooking process. As EP09 puts it: " I found it now neither annoying nor that I was too much restricted, because I mean at the end of the day I can still vary it myself". The *creative experts* point to the value of the assistant for other users, but refuse to use it in their daily life, preferring to do the tasks themselves, with the exception of EP02, who considers herself "prone to chaos". Also noteworthy was that participants within this group expressed a high interest in background information.

As in the observational study, only one participant was classified as *spontaneous creative* (EP06), characterized by a low cooking competence and a high need for autonomy in cooking. According to the pre-study results, such users are the most difficult to support, because the primary goal is to recognize potential mistakes. In line with a high need for autonomy, EP06 often questioned the decisions recommended by the assistant and would have appreciated more background information and explanations. Further, she cherished the opportunity to make choices at the decision points and would have liked even more of these options.

### 8.7.2 Automation & Autonomous Events

Overall, participants responded positively to the topic of assistance functionality based on proactive events. Among others, they mentioned that the monitoring of device data could prevent users from making mistakes. For example, the prototype uses data from the Cookit's weighing scale to inform the user about forgotten ingredients. Some participants did not realize that they were supposed to add sugar while stirring the egg whites until the assistant reminded them. In the post-interviews, the participants recalled a number of situations when distracted, stressed or forgetful, and being grateful for a reminder, as e.g. EP06 notes: "You don't end up standing there thinking like, 'Crap, I forgot the raisins.' " Another example is preheating the oven. If the user forgets to turn on the oven in the beginning, by the time the user realizes the problem, it will add on the total time of cooking.

Additionly, participants expressed appreciation for the facilitation of work that smart assistance provides, as EP03 states:

> "The first time, 'wow', I said and really felt that deeply 'wow', that was when he [Cookie] [told] 'I now turn on the oven for you, you do not need to worry about it. It'll be warm when you need it.' "

Another aspect of facilitation is that users do not have to look up instructions (again) when settings are transferred automatically. EP07 alludes to feeling flattered that the assistant is doing the work for her: "I don't have to worry about it at all? Ey, I think that's great." [...] Yes, I feel honored." However, while speed and convenience are increased, automation may preclude the possibility of users learning to do the tasks themselves. As EP08 puts it: "I also had the feeling the whole time that I don't want to press anything in between. Perhaps simply because I have never operated the stove manually and I don't really know how to operate it manually."

### 8.7.3 Feeling of Control

During the development of the prototype, we intentionally refrained from asking the user for authorization for every automatic appliance setting made by the assistant. The idea was to observe how participants would react to relinquishing control and relying on the assistant's competence. Consequently, participants indicated that they would prefer to be consulted most of the time before the assistant automatically adjusts settings. In some cases, the participants even felt externally controlled by the assistant. This was voiced explicitly, e.g. by EP09: "I still find it a bit strange, because you are somehow controlled by something else to do the whole thing" but also observed when the assistant reacted differently than the users expected. To counteract the users' feeling of loss of control, a detailed and accurate intention recognition is needed. When the user's intentions are clear and the interpretation of the user's actions is correct, the assistant's autonomous and proactive support is aligned to the user's needs and no longer surprises or interrupts the user.

As soon as a system performs actions autonomously, without the explicit consent of the user, safety becomes crucial. For instance, participants described cases where a potentially flammable object was left on the stove while it was turned on. It was noticeable that there were clear differences between the different appliances regarding the participants' perception of proactivity. While all participants appreciated the fact that the oven was turned on and off autonomously, the opposite was true for the stove's autonomous settings. Participants explained the difference by saying that the stove demands more active operation and control with the user in front of it, whereas the oven is more remote and requires less interaction or attention. Very similar to EP03 and EP05, EP06 summed it up as follows:

> "No, with the oven, I thought it was good. It's far away enough. I think because with the oven I don't have to do anything actively myself. It's good when it just does everything, especially when preheating, I really don't have to do anything."

### 8.7.4 Transparency & Trust

As discussed in Section 8.5, the assistant prototype "Cookie" was designed to communicate its actions transparently, e.g. "I'm going to preheat the oven for you." In turn, participants responded that this type of communication reinforced their confidence in the settings by the assistant, e.g.:

> "And these indications that she has now basically also turned on the stove top or just I'll say now, that then activated accordingly. 'So I noticed, you did this and that.' I thought that was good because, with the background that if you completely trust it, then it's such an interaction. Well then, then you don't feel so lost." (EP09)

Some statements of the evaluation demonstrate that the trust of the participants increased once the causes and reasons of an autonomous event were explained in greater depth. For instance, the assistant informs the user that it has detected the pan being placed on the stove and will now turn on the oven to prepare the appliances for the upcoming step. However, "Cookie" did not provide a central dashboard for the status of all appliances in the kitchen due to a lack of technical integration. Nevertheless, it is fair to assume that such a dashboard, displaying running timers, would further promote trust in intelligent assistants that proactively perform tasks for users.

### 8.7.5 Multimodal Interaction

Our findings suggest that user output and input ideally require different interaction modalities to support the user in an efficient and desirable way. While participants used the voice interface mainly for input, they preferred visual representations for the output. Six out of ten participants mainly or exclusively used voice as input modality, while the other four participants combined voice and touch input during the experiment. One of the several reasons for preferring voice input is the use case of cooking: Users prefer voice over touch control because their hands are usually busy and dirty from cooking. In these situations, voice control simplifies the process by reducing the need to wash hands before interacting with the assistant. Additionally, voice control is time-saving. In most cases, users are still engaged in a cooking activity while looking for the next step or extra hints. Voice control enables the user to take control from their current location in the kitchen, rather than interrupting the ongoing activity by moving to the touch interface. However, some participants have concerns about using voice as an input modality. EP09 wonders if the assistant would really understand everything at the first attempt. Additionally, EP06 states that using a wake word can sometimes be inconvenient. Although all participants preferred voice as an input modality, they did not use it exclusively. Instead, they opted for touch when it seemed more efficient to them, for example, when standing directly in front of the tablet. Instead, they opted for touch whenever it seemed more efficient, such as when standing directly in front of the tablet, as EP08 stressed: "If I can still control it with my fingers, it's actually always better to do that because it's always faster. Especially when there is already a button for what you want to do." EP03 remarks: "Sometimes I'm already there anyway, [...] it's faster for me to click than to articulate anything now".

Unlike the input interaction, the examination of the output interaction reveals that the participants tend to focus more on the graphical and visual interface. This is partly due to participant retention. Cooking activities create noise that makes it difficult to hear and understand what the voice assistants are saying. Therefore, participants emphasized the availability of a visual representation of the recipe to review the necessary information or when processing new information. Hence, key information should always made visible. The different forms of presentation, such as videos, images, or text, help to visualize what the recipe outcome should look like and how it can be achieved. Several participants mentioned that the pictures and videos provided useful orientation, e.g. EP02: "Then you can always compare how it should look like." When no visual representation was available, participants were sometimes uncertain whether the result was correct. Overall, the combination of the different output modalities is perceived as highly intuitive and used more likely than just one output format.

Furthermore, participants should constantly be able to track the progress of the recipe. This in itself created a sense of security and facilitated the linking of recipe steps. Particularly, the overview diagram was mentioned as beneficial for communicating this type of information. It shows exactly the number of recipe steps, their relation to each other and the current progress at defined milestones of the cooking process. Besides, clicking on the overview icon or using the voice control made it permanently available and easy to retrieve.

### 8.7.6 Personification of the Assistant

Every dialog-based system has a characteristic conversational component, which in the design of "Cookie" covers the feature of guided cooking. Our findings reveal that participants appreciated the human-likeness of the system. Some participants explicitly voiced the impression of someone cooking with them along, e.g. “I feel like I’m not cooking alone” (EP07) and “This feeling, hey, there’s somebody talking to me, it’s not a beep, it’s not a honk or something, it’s the spoken word; I think that’s good.” (EP07). Furthermore, the conversational component makes it possible to praise and validate the user. This can promote a positive and motivating atmosphere for the cooking process. Not surprisingly, some participants refer to Cookie as if to a person, e.g. EP06: “I don’t remember exactly what she said - but she always gave praise.” And EP02 remarks she feels “pleased, even if that’s a device that praises you”.

### 8.7.7 Summary of Findings

In this section, we present an overview of the evaluations’ main findings as a basis for the discussion.

**8.7.7.1 Users in Control** Analyzing the results according to previously categorized cooking types revealed that participants appreciated the assistant adjusting some settings autonomously and engaged in taking control of decision-making as it was relevant to them. Particularly, prevention of mistakes, alternative action paths, alternative modes of instructions, and the design of background information promoted a positive and pleasant experience of decision autonomy. Participants expected advice to be successful without compromising taste. But they wanted more freedom when making less critical decisions. Likewise, they disliked the feeling of external control and demanded prior consultation before automated settings were made. This behavior contributes to a successful expectation management for users. Consequently, every appliance requires different levels of user control or automation.

**8.7.7.2 Trusted Automation** Participants perceived the facilitation of work processes that reduced mental effort as appropriate support, e.g., when automating small actions like turning on the oven. Further, they appreciated the reduced need of repeatedly looking up instructions and the prevention of mistakes due to reminders. Different modalities to communicate information enhanced the trust building between participants and the assistant, e.g., participants valued being shown confirming information like timers and explanations for automated actions. Over time, users will require less explanations due to their increased trust in autonomous systems based on fulfilled expectations. Finally, future work should provide a dashboard for an activity overview of the kitchen assistant and the orchestrated appliances.

**8.7.7.3 Multimodal and Personal Assistant** The employed combination of visual and voice interaction modalities was well received and proved to optimally support cooking activities. Overall, all participants preferred voice input as they considered this interaction mode as practical, convenient, ubiquitous, and time-saving. One exception was a preference for touch input in cases where a single button click would have been sufficient. Participants emphasized the importance of visual displays as cooking noises make listening comprehension difficult. Further, visualizing the overview and single steps of the recipe supported understanding the cooking process and reaching the desired outcome. The assistant's multimodal and speech capabilities contributed to the perception of the assistant being as a human-like and personified artifact, even without using an avatar. The assistant cooking along and giving occasional praise and validation contributed to a motivating, less lonesome and entertaining atmosphere.

## 8.8 Discussion

As outlined in Section 8.2, the presented studies were motivated by the question of how an integrated, multimodal and proactive system needs to be designed to assist users beyond automation, honoring their desire for competence and autonomy and supporting the specific (food) practices and challenges while cooking. Thus, this work adapted general design principles for interactive systems [365] to the specifics of cooking to enable studying the ramifications of the resulting (prototype) system in-depth with potential users. The Wizard-of-Oz methodology employed in the evaluation facilitated a focus on qualitative user research instead of software implementation. Based on our observations, we derive the following implications:

### 8.8.1 Pursuing Matching Levels of Proactivity

Based on the observational study and related work (e.g. [365]), we hypothesized that the provision of an appropriate level of proactivity depends on the characterization of the users' personality and an accurate assessment of their current situation. Indeed, the categorization of the four user types derived from the observational study proved to be instrumental, since these different user groups exhibited different needs of trust, autonomy, and competence to take advice from the assistant. We highlighted this in our user study with "Cookie".

Users classified as beginners closely followed Cookie's suggestions and instructions and perceived the provided information on choices and background knowledge as an opportunity to acquire know-how. In contrast, agency is conveyed to *accurate cooks* by the opportunity to explore alternative paths and choices, and potentially, receive support with more advanced recipes and areas they are less familiar with. For the more "creative" cooks (creative experts and spontaneous creative) with high demand for autonomy, the role of proactivity is emphasized for preventing mistakes. The need for autonomy is expressed by the observation that these users want to retain control, by confirming proactive offers or being able to turn them off.

Related work [173] has previously found that moderate levels of proactive intervention, e.g., notifications or suggestions, elicit higher levels of trust in users than autonomous intervention. In our design, reminders and suggestions fall into the former category, and users in our study actually received them well. However, users experienced a loss of agency as a result of autonomous events that did not wait for user consent when "Cookie" operated appliances autonomously, especially in situations considered safety-critical or intricate, for example when operating the stove, as opposed to preheating the oven. These findings have practical design implications for the equipment of proactive cooking assistants. Users should have an explicit choice and setting option for each appliance and function to be controlled autonomously by the system or manually by themselves. Thus, our study supports the findings by Kraus et al. but also emphasizes that accurate intention recognition and transparent explanations are crucial to underpin proactive system behavior. Although future work will be needed to develop a robust instrument to use generalized cooking types for technology development, we made a first step to sensitize for valuable differences in proactive design. Our Wizard-of-Oz setting allowed us to provide the participants with a system behavior based on an accurate assessment of their intentions. Since the wizard interpreted the spoken requests to the assistant while fully observing the participant, “Cookie” provided ideal conditions for intention recognition. It should be anticipated that shortcomings of a fully automated system in terms of speech recognition and detection of user activity may lead to less trust than that expressed by the participants in our study, making it even more important that users are enabled to exert control on the degree of automation.

### 8.8.2 Adaptive Solutions Contribute to Human Agency

Our observation showed that users vary in their need for autonomy and have different levels of cooking competence. That impacts their perception of their agency and the assistant equally. Although our evaluation shows that most of the participants felt controlled by the agent, all of them praised the experienced support provided by the assistant. So far, automation is helping to ensure consistent meal preparation results in fast food processing [217]. This can be extremely valuable when cooking in a stressful setting or under time constraints, in general. Cooking for guests or loved ones could also demand for reliable results. Our approach shows that users are eager to learn more background information concerning food preparation, as postulated by Kato et al. [157] and respectively tool handling for cooking. As our findings show, the automatic on/off of the cooktop deemed to be useful in case of danger. Thinking further, cooking assistance could also promote a better understanding of heat control as an integral practice in meal preparation, e.g., different dishes require appropriate levels of heat. This kind of automation is directed to teach the influence of the (connected) thing respectively cooktop and might positively change humans’ food practices in the future [3]. By repeating the proper cooking practices as a co-performance of the assistant and users,

users' skill level can increase and they learn the consequences of their actions. Consequently, as the practical application of knowledge promotes the user's learning of cooking methods, their immediate need for assistance for mastered recipes decreases. Therefore, the assistant should adapt by introducing new methods, ingredients as variations or related recipes to advance user skills further. However, our study results cannot report on any long-term effects on society. Nevertheless, we could observe that a balance of agency contributes to a positive attitude towards technology and learning of new skills. However, we need further research and design for food practices that promote human agency over automation, as suggested by previous research [3, 79].

Dolejšová & Wilde et al. [79] have cautioned that overemphasizing automation and technology-centered designs in general might distance users from their socio-cultural meaning and material of food practices. As our findings suggest, not all participants embraced all recommendations by the assistant. Instead they individually and deliberately decided which risks to take or which kind of personal variation to follow. Those participants with a high need for autonomy from the beginning benefit either from error prevention or, for those with an additional high level of competence, from challenges that contribute to more creative approaches. The exit points we planned and implemented during the process were also well received as support for autonomy. By not restricting the order of the recipes in the evaluation, one participant found that the suggested order was reasonable and saw his alternative decision as a learning opportunity. As our users felt somehow under control by the assistant during the evaluation, the assistant might follow a conservative approach in support not to risk creating tensions [57]. Future versions of adaptability and flexibility should offer an over-write function to save personal variations that could be further shared with family members and so on. This would underline the meaning of recipes and creativity and self-expression in cooking rather than aiming for a "perfect" but standardized version [79]. In any case, adaptive solutions should aim not to constrain and control the situation too much, but to adapt to the personalities and context of the users. By reducing IoT interventions, users are given the opportunity to reclaim their autonomy and feel empowered by cooking for themselves and others.

Although we did not pursue an anthropomorphic design with our assistant, participants attributed a positive and motivating atmosphere to the assistant. They appreciated the social presence [267, 164] the voice embodied and not having to cook alone. As a result, we can assume that prospective designs might explore how the personalities and designs of such assistants can positively impact the pleasure of cooking [5]. Further, the assistant might also motivate users to engage in new hobbies and reinforce their agency. However, we can only speculate cautiously, as we used a Wizard-of-Oz approach and could not completely rule out human factors. Nevertheless, we see potential scope for design to further strengthen human food practices and encourage activities that users enjoy doing.

### 8.8.3 Multimodal Interaction for Interactive Support

The results of the user study also indicate that user perception of information can vary greatly from case to case. This depends, for example, on the type of output modality, but also on the information presentation within a single modality. For example, via acoustic feedback from the system, a beeping sound may be less meaningful than a voice output that articulates clear suggestions or warnings to the user. During the design process, information needs to be prioritized and ideally classified according to its salience like, e.g., the prioritization levels for hints as discussed in Section 8.5.2, and how these salience levels are communicated to the user.

Based on our findings and research [345], we propose that multimodal interaction can compensate for the weaknesses of single modalities, examples of which are the limitations of voice interaction in a noisy kitchen environment, the limitations of touch when users' hands are dirty, or the difficulties of gesture recognition, which require training on the part of the users. Furthermore, the use of multiple modalities promotes proactivity, offering multiple ways to intervene with varying degrees of intrusiveness, e.g., providing hints by voice vs. on screen, or both. Accordingly, "Cookie" has been designed to appropriately combine GUI and voice interaction, capitalizing on users' clear preference for voice over gestures; in particular to offer voice as an input modality that is backed up by GUI input, and a combination of voice and GUI for output. While the merits of gesture interaction became clear from the analysis of the observational study results, gesture interaction was not integrated into our design of "Cookie". This avoided potential inaccuracies in recognition, such as those reported in studies with Kognichef, cf. [229], as well as the necessary instruction of participants in the use of gesture control, which can be considered time-consuming and intricate on its own. The participants' assessment confirms the viability of the chosen combination of modalities for "Cookie", but does not preclude potential improvements obtained by the inclusion of gestures or connected/instrumented cooking utensils (cf. [303, 251]) into the interaction concept as part of possible future work. This would provide an opportunity to (further) incorporate cooking knowledge embodied in the handling of kitchen and food items into the assistance concept, as suggested, for instance, by Baurley et al. [14].

## 8.9 Conclusion

Based on an in-depth observational study focusing on users' assistance needs, this research aimed to design a holistic interaction concept for assisted cooking in smart kitchens. We derived several design considerations regarding appropriate interaction modalities and proactive user support. Our results outline how to combine multimodal interaction styles to reduce the occasional intrusiveness of proactivity, while considering users' needs for autonomy and competence in the design of proactive events and behaviors. In particular, error prevention and work facilitation were perceived as positive. However, this requires an accurate interpretation of users' intentions and situational context. The technological challenges to manage and extend a corpus of recipes and associated multimodal instruction materials remain as a topic for related and future work. This research highlights the importance of transparency in the integration of multiple connected appliances. It also calls for a balanced autonomy continuum to provide users with a greater sense of control. Further research is necessary to determine the detailed effects of user personality on their perception of decision autonomy as well as trust in proactive system takeover.

# 9 "Foggy sounds like nothing" - Enriching the experience of voice assistants with sonic overlays

## Abstract

Although Voice Assistants are ubiquitously available for some years now, the interaction is still monotonous and utilitarian. Sound design offers conceptual and methodological research to design auditive interfaces. Our work aims to complement and supplement voice interaction with *sonic overlays* to enrich the user experience. Therefore, we followed a user-centered design process to develop a sound library for weather forecasts based on empirical results from a user survey of associative mapping. After analyzing the data, we created audio clips for seven weather conditions and evaluated the perceived combination of sound and speech with 15 participants in an interview study. Our findings show that supplementing speech with soundscapes is a promising concept that communicates information and induces emotions with a positive affect for the user experience of Voice Assistants. Besides a novel design approach and a collection of sound overlays, we provide four design implications to support voice interaction designers.

## 9.1 Introduction

For several years now, households talk to Voice Assistants (VAs) in their homes and welcomed them as everyday companions [206, 254, 20, 60]. Usually, most users use them predominantly to control and access home appliances and internet-based services [6, 289, 206], e.g., playing music, setting alarms, requesting weather forecasts, or asking for specific information [6]. By now, VAs have a significant contribution to the consumption of and interaction with information [206].

The progress in speech synthesis [1, 169, 288, 286] and voice design [284] allows to make voices more human-like [284], less annoying [73], more appealing [39], more charismatic [347], or provide contextual cues implicitly [284]. In addition, the new opportunities offer designers to play with gender stereotypes [312], enable voice branding [162], or enrich the voice experience in general [168].

However, most users expect efficient and convenient interaction in a utilitarian sense as past experiences have disappointed them due to a lack of personal bonding and emotions [53]. Apart from considering the voice interaction as boring and monotone [243], users hope for a lively assistant, resembling a friend, that can express opinions and emotions itself as well as engage in a conversation [53].

In addition, the auditive channel bears potential in making use of sound design. Several researchers propose to explore interaction and experience beyond the dichotomie of human and machine and establish new design approaches for voice interaction [60]. Meanwhile, further researchers emphasize to integrate more sound design as well [298, 49]. The principles of sound design as there are sonification of data and interactions [271], musical expressions [152], the design of earcons and auditory icons [27, 107] represent great potential to enrich and enhance the current state of VAs. As stressed by Fagerlöhn and Liljedahl: "Sound design can be described as an inherently complex task, demanding the designer to understand, master and balance technology, human perception, aesthetics and semiotics." [191].

While sound and the sonification of data could supplement the repertoire of speech synthesis and voice design by communicating information and expressing [271, 152], e.g., moods, atmospheres, emotions, interaction designers have not systematically adopted these extra options, so far. In this light, we draw from concepts and theories of sound design to explore our following two research questions:

**RQ1** How might sound add to the user experience of Voice Assistants?

**RQ2** How can we use sonification of data in information design for Voice Assistants?

In our work, however, we consider sound in its serving function to illustrate and enrich what is spoken by a Voice Assistant. In other words, we focus on the overlay quality of sound as a supplement to the speech output. As weather forecasts are a frequently used service of VAs, we decided to investigate this use case and its sonification. Therefore, we first conducted a user survey with 33 participants to empirically gather associative concepts and sounds for seven perceptible weather conditions. In the next step, we analyzed the design material and developed a sound library of seven distinct audio clips that illustrate our concept of sonic overlays. Finally, we evaluated and discussed our library with 15 participants in a qualitative interview study.

Our work shows that complementing voice interaction with illustrative soundscapes enriches the communication of VAs and is appreciated by potential users. As our empirical findings reveal, layering sound and speech needs special consideration of the relation of both and in light of the intended message. Therefore, we propose a user-centered design approach grounded in sound design that employs conceptual associations and the combination of iconic, abstract, and symbolic sounds. Sound Overlays, as outlined in this paper, could be used as an alternative to the advancements in speech science that focus on the modulation of emotions through the use of voice and speech as a design material. Furthermore, implementing a sound design in voice interaction might complement the emotional tone of voice of VAs in future designs. Soundscapes in voice interaction design add to the atmosphere of speakers to tell thrilling stories, as we know from sound design practices of modern media. Finally, we propose four design implications: Investigating soundscapes for voice interaction design (1), supplementing vocal messages by sound (2), aiming for authentic soundscapes (3), and finding a balance between expressiveness and informativeness as well as coping with trade-offs between clarity and sonification of information (4).

## 9.2 Related work

Our work is grounded in the following research fields in particular: VAs and Voice Interaction Design (see Section 9.2.1), earcons and sonic information design (see Section 9.2.2), and the design of sound effects and for sonic experiences in general (see Section 9.2.3). The first field focuses especially on the use of speech to enable natural conversational interaction with the user and addresses advancements in speech sciences to reflect on vocal speech as a key design material in voice interaction design. The second field deals with the auditory sense as an additional channel to encode and convey information. In terms of this work, we understand *encoding* of information as the process of using auditory channels to express information that humans can process with their auditory senses and understand in a meaningful way. Contrasting to the previous perspectives, the latter focuses on the effect and use of soundscapes in related fields of HCI and investigates the use of sound effects to enrich the experience of interactive media. To the best of our knowledge, only a few studies adopt concepts from sound design in the context of voice assistance and voice interaction design. In particular, current voice interaction research focuses on speech exclusively to make the voice output more natural and informative.

### 9.2.1 Voice interaction design

Voice interaction design represents a new type of interaction [243] that is primarily concerned with encoding and conveying information in spoken language. Particularly, the text-to-speech capabilities of current Natural Language Processing (NLP) machines [227, 313] enable and drive this emergence and growth of voice-first applications. The ephemeral character of speech-embodied information in comparison to text reveals different challenges of information communication by VAs, such as cognitive load or dead end conversations [60, 254, 298]. Due to a lacking persistent manifestation, cognitive load is increased and listeners are required to deeply focus in order to process and react to information [298]. Grice [97] argues that communication practices should always consider the quantity (right amount) and quality (speaking the truth) of information, as well as sharing only relevant information with a maximum of clarity.

However, user expectations regarding the capabilities of VAs remain frequently unfulfilled and cause disappointment and frustration as they expect an effortless and engaging exchange of information [53, 89]. Often, well-known usability issues like limited NLP and speech recognition, system errors, misunderstandings, and failed feedback cause this phenomenon [221, 289, 68]. As a result, this leads to an interaction style that is based on “guessing and exploration [rather] than knowledge recall or visual aids” [221]. Additionally, this type of conversational interaction does not feel natural, and lacks sufficient positive experiences to motivate users to engage frequently [53]. Consequently, VAs need reliable usability to prevent users from negative experience [60, 221, 68], and furthermore research to investigate the positive aspects of user experience, which might contribute to an enchanting, playful, meaningful, and engaging interaction [38].

Accordingly, current research studies anthropomorphic effects and how to mimic human-human conversation successfully [9, 68], even though some research points to negative effects of too much human likeness [68]. Further experience dimensions for conversational agents might build on a more flexible attitude regarding the categories of "human" and "machine" [60, 130] and [20, 243, 289] should "fit into and around conversations" [243], and respectively routines of the users. We should understand speech as an act of performance, a kind of storytelling [9], and affective communication strategies [197] to enrich the interaction and stimulate experiences. For instance, new modes of articulation like "whispering" already extend the dimension of sonic experiences and prevent the VA from being perceived as boring and monotone [243].

Moreover, human information processing is not linear but complex. The Elaboration Likelihood Model [248, 249], for instance, stresses that humans process information via two routes: via the central route, people decode the content of the message by listening carefully to the semantics, the strength of arguments, and the credibility of included facts. In contrast, via the peripheral route, people respond emotionally to the message, where they are more likely to rely on general impressions, peripheral cues, and subliminal tones.

Affective and emotional speech research [198, 347], especially speech emotion recognition [1, 169, 288], emotional speech synthesis [286] and emotional speech production [187] represent an emerging research area addressing these subtle but vital aspects of communication. A body of work studies, for instance, how our voice and our way of speaking express a range of emotions like sadness, joy, anger, dearness, surprise, and boredom [153, 284, 286]. Furthermore, various studies have shown that speech and voice impact credibility, trust, charisma, attractiveness, likeability, and personality perception in general [284, 290, 347].

Research, machine learning in particular, also underlines the features responsible for communicating emotions. For example, research on emotional speech uncovered that acoustic levels such as frequency, bandwidth, pitching, intensity, loudness, speech rate, pausing, duration, and intonation of phonemes, words, and utterances influence the perception of emotions [1, 198, 287, 288]. Further, several linguistic and paralinguistic, among other more abstract features like gender, age, or accent, influence users' perception of speech and voices [1, 284].

Regarding speech emotion design, researchers have specified various notation systems, such as the emotional markup language [37, 87], which allows designers to annotate parts of sentences to be spoken with a particular emotion. To support designers, Shi et al. [294] outline the concept of state-emotion mapping that may serve to drive human-VA conversational interaction. However, to save designers this additional annotation work, the researchers proposed a text-based emotion detection algorithm to contextually determine the emotional phrasing and pronunciation of sentences [198].

Our approach aims to supplement advances in speech science that focus on modulating emotions through speech to create engaging experiences between users and VAs by investigating alternative interaction design approaches.

### 9.2.2 Sound design and data sonification

Even though sound design is an active research field in the HCI community, there is a call for more scientific approaches to enable reproducible results [191]. So far, this field moves between craftsmanship and art and depends on skillful sound designers, as "Sound design can be described as an inherently complex task, demanding the designer to understand, master and balance technology, human perception, aesthetics and semiotics." [191]. Sound is an integral part of media and system design to convey a captivating narrative, and an integral component for audiovisual storytelling [276].

Therefore data sonification represents an integral process to encode data and interactions so that the intended meaning is not misunderstood. According to Enge [88], sonification can be seen as "the use of nonspeech audio to convey information" [201], whereas visualization is understood as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" [228]. Visualizations support a clear understanding of information, while sonification frequently allows for more interpretation despite its means to convey information [271]. Therefore, the most common approaches to auditorily encode information in interaction design are auditory icons and earcons [27]. A fundamental difference between auditory icons and earcons is that earcons can be considered to be arbitrary symbolic representations, while auditory icons can be regarded as analogical representations. Blattner et al. [27] defined earcons as "non-verbal audio messages used in the user-computer-interface to provide information to the user about some computer object, operation, or interaction". Brewster further specifies that earcons are "abstract, synthetic tones that can be used in structured combinations to create auditory messages" [34].

The sonification of data is not only able to encode information but is also capable of expressing and inducing emotions. Depending on the design goal, inaccuracies may exist, as humans evaluate emotions very subjectively [271, 152]. Thereby, experiences are based on the affective and functional perception of the design. This poses a challenge to research since it aims to investigate sonic elements and their impact objectively but competes with the narrative qualities of music and its affective and emotional impact [292]. While an interesting and positive experiential design may stimulate emotions, there will be a trade-off between the sonic experience and the clarity of the information [271]. The expression of emotions is defined by its psychophysical relationships between musical elements and perceptual impressions of the user. Further, capturing emotional expression in music is possible by focusing on a listener's agreement as no one can effectively deny their experience [152, 41, 80]. In contrast to expression, communication further depends on accurately recognizing the intended information and emotion [152, 154]. Therefore, our work aims to explore the relation between a clear understanding of information and the enrichment of emotions by combining sound and speech.

### 9.2.3 The role of sound design in modern immersive media

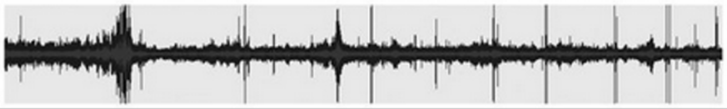
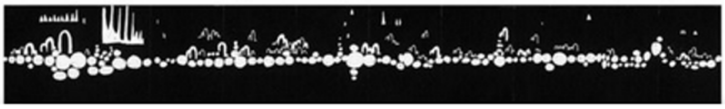
Following Simpson [298] and Sanchez-Chavez et al. [49], scholars argue that advanced methodologies and design principles for Conversational User Interfaces (CUI), e.g. interfaces for VAs, chatbots, are needed. So far, current designs follow engrained and trusted GUI principles to present and represent information without considering the dimensions of auditive information processing, for example, the ephemeral state of speech, memory, imagination, user interpretation [298, 49, 60]. Sanchez et al. [49] propose to even go beyond current conversational design "to include more nonverbal and paralinguistic elements" that could expand the design space further when considering sound interaction as a primary form of interaction.

In the light of the above, in most cases, sound is regarded as a complementary approach to enrich the experience of visual media like in games, and movies: "Auditory cues play a crucial role in everyday life as well as VR, including adding awareness of surroundings, adding emotional impact, cuing visual attention, conveying a variety of complex information without taxing the visual system, and providing unique cues that cannot be perceived through other sensory systems. [...] VR without sound is equivalent to making someone deaf without the benefit of having years of experience in learning to cope without hearing" [150]. Further design studies revealed that soundscapes effect tasting experiences by adding a significant hedonic value [46, 341]. Soundscapes are defined as an "acoustic environment as perceived or experienced and/or understood by a person or people, in context" [143], which means that they represent a sign to their perceivers. We can also observe that, for example, conscious choosing of sounds plays out differently in behavior stimulation of children regarding play experience and the play itself [140]. Overall, sound design creates imaginative spaces in research and practice and is particularly important for narrative designs [56]. Adopting sound design principles for voice interaction design, we aim to enrich the narrative strength of VAs and explore how this will affect potential users.

## 9.3 Conceptualization and empirical investigation of a sound library

Weather reports are a frequently used service of VAs by users. In light of our research questions, we aim to build and evaluate a library of sonificated weather reports as a case study. Thereby, we decided to adopt the approach proposed by Mynatt et al. [224], who discussed potential pitfalls during design and subsequent recognition failures by users during the use of a sound-based interface in their work. In particular, the authors emphasized considering four categories for designing auditory icons: identifiability, conceptual mapping, physical parameters, and user preference. As follows, we discuss relevant theoretical concepts from

**Table 10**
Enhancing the voice track with a sonic overlay

| **Speech Output** | the weather in cologne on monday is sunny |
|---|---|
| 1st Track **Voice** |  |
| 2nd Track **Sonic Overlay** |  |

related fields of sound design. Second, we continue with a user survey to collect conceptual mappings and physical parameters as design materials to empirically ground the design space for sonic overlays.

### 9.3.1 Theoretical implications from sound design

Current design practices of VAs focus on advances in speech modulation and interaction while not having established to complement speech-based output with soundscapes, yet. In this context, a sonic overlay can technically be characterized as a second track played in parallel with the voice as the primary track (see Table 10).

Regarding the goal of sonic overlays, two fundamental requirements can be identified that the design should take in mind:

- Discrimination quality: As the primary information is given by speech, the sonic overlay must not impede or interfere with the information transmission of the first (talkative) channel.
- Conceptual mapping: The second track is not arbitrary but should supplement the first to render the output more expressive and informative.

**9.3.1.1 Increasing the discrimination quality of sonic overlays** In contrast to earcons, the aim of sonic overlays is not to substitute and summarize one specific piece of information but to enhance the experiential quality of information articulated via speech. Therefore, voice and sonic overlays have to be designed in synchronized co-existence to communicate and express information auditorily and in parallel. Hence, we take a special focus on what we call the discrimination quality – a category and feature that allows the user to isolate, separate, and process speech- and sound-based information directly.

Krygier [176] has adopted the basic concept of visual signifiers to the auditive channel. He outlines the concept of sonic variables by focusing on abstract sounds that can be modulated

**Table 11**
Sonic variables and their discrimination quality related to voice output.

| Variable | Description | Discrimination Quality |
|---|---|---|
| Spatial Location | The location of the sonic overlay related to the voice output in a two- or three-dimensional space | Effective, depending on the location distance |
| Loudness | The magnitude of the sonic overlay are related to the voice output | Effective, depending on the volume distance |
| Pitch/ Frequency band | The pitch of the primary frequency band of the sonic overlay is related to the voice output frequency band | Most effective when the frequency band of the sonic overlay is below or above the voice band (a partial overlapping is possible) |
| Timbre/ Sound Motives | The general prevailing quality or characteristic of the sonic overlay related to the voice output | Effective when the timbre of the sonic overlay is different from the human voice (e.g. music, abstract sounds, or natural noises) |
| Temporal position | The temporal location of the sonic overlay is related to the voice output | Most effective when the location is before (intro position) or after the voice (outro position. A partial overlapping and fading is possible) |

by frequency, volume, or timbre to encode information. Studying the variation systematically, he concludes that sound location and volume, pitch, register, timbre, duration, rate of change, order (sequential), and attack/decay are viable sonic variables to enhance geographic visualization. In contrast to Krygier [176], we move the design space beyond abstract sound and consider speech-based output as embedded and discriminable quality of a holistic audio clip. In this sense, Table 11 presents a not conclusive set of sonic variables that aims at the most notable discrimination possible between sonic overlays and speech-based output.

For our design, we took the discrimination variables *Loudness*, *Timbre/Motives*, and *Temporal position* into account which we regard as most impactful in our design. We discarded the variable *Frequency band* because we aimed for simple and non-modified soundscapes. As smart speakers vary in their technical loudspeaker quality, we neglected to build on *Location* as discriminative quality. However, this dimension might be worth considered in future design studies, as certain listeners using high-end speakers and headphones for VAs on their smartphone, have the technical equipment to experience localization in 3D sound spaces. It might support immersion by, for example, indicating the incoming direction of wind and rain in acoustic weather forecasts. In the following paragraphs, we provide further detail to understand how the chosen variables add to and are reflected in our design.

**Loudness** Humans can distinguish between different volumes from about 3 dB up to 100 dB. Loudness owns an ordering function by its nature. Keeping a sound experience linear without any variance, loudness might become unconscious over time. Hence, different magnitudes of loudness might highlight and contrast parts of the sonic experience [176]. In particular, different volume levels might increase the discrimination between speech- and sound-based information by lowering the illustrative sounds and turning up the voice volume.

**Timbre/motives** Krygier [176] defines timbre of sound as the encoding of information by the character of a sound. In analogy, instruments own a characteristic sound, such as the brassy sound of a trumpet, the warm sound of a cello, or the bright sound of a flute. Similar to the human voice, Alexa, Siri, and other VAs have a distinct sound that is distinguishable by the human ear. By choosing and incorporating distinct timbres for sonic overlays, their discrimination quality might be increased. Consequently, using tones or pieces of music, like a bird's flutter or a synthetically produced ambient sound, contribute to recognizing both auditive tracks. This way, information on both tracks can be encoded independently. Additionally, music and sounds transport atmospheres and expressions of emotions, often recognizable as a distinct motive and in movies even underlining principal characters. Such superimposition of motives supports the construction of compound earcons [27] but can also be applied to sonic overlays.

**Temporal position** By its very nature, audio tracks have a temporal structure and order. Thus, discrimination can also be supported by separating the sonic overlay and the voice track in time. The intro and the outro take a particular temporal position here. For instance, either speech may start or the sound of falling raindrops before the assistant begins talking. Further, incorporated background sounds may support the discrimination of auditive information when speakers pause.

#### 9.3.1.2 Conceptual mapping: the semiotic of sonic overlays

We aim to create sonic overlays that are not arbitrary but related to speech-based information. The main goal of sonic overlays is to serve as an illustration of what has been said, leading to double encoded information by speech and a sonic overlay. For instance, if the VA reports rain for the next day, the sound of heavy rain supports this information. To characterize the relation between the vocal output and the sonic overlay, we apply Peirce's semiotics [244] similar to David Oswald [237] in his work about the semiotic structure of earcons. The core of Peirce's semiotic is the symbol as a triadic relation between the object, the interpretant, and the sign:

- Sign: the sign-carrier which has a perceptual representation
- Object: a thing, a concept, an experience, or an emotion the sign refers to.
- Interpretant: the perception and interpretation in form of perceived object mood, or emotion in the mind of the perceiver

The sign mediates between the object and the interpretant. For instance, the ringtone of the mobile phone mediates its owner that someone is calling her. In this case, the knowledge of the calling is the interpretant, and the referred call presents the object, while the ring tone is the sign that caused that interpretation. In Peirce's semiotic [244], we can say that the linking of the mobile phone's ringing and its vibration refers to the same object (the call) as well as the interpretant (the knowledge of the call). In the same way, we can now characterize the relationship between the speech and its sonic overlay.

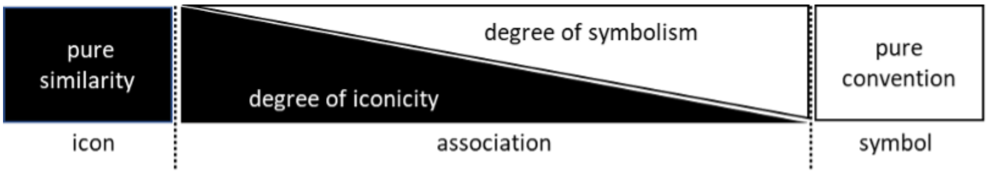


**Figure 11**
Gradual transition of icon to symbol, from high iconicity to high conventionality (adopted from [237])

Looking at the encoded meaning in this process of creating sonic overlays [237], Gaver[107], for instance, distinguishes between an iconic, a metaphorical, and a symbolic perceptual mapping. In contrast, Oswald [237] uses the Peircean tradition [244] distinguishing between iconic, indexical, and symbolic signs. Our view is influenced by both authors. Focusing on the experience, we follow Oswald's comment that the constitutive element for iconic signs is similarity, not physical causality. For the same reason, we focus on associations, metaphors, and signal correlations that establish a link between a sign and its object. Consequently, we distinguish between three sign categories referring to the three kinds of relationships:

- Iconic: the representation based on the similarity of the signs and the signals produced by the object
- Associative: the representation based on associations, metaphors, or correlations between sign and object and the signals produced by the object
- Symbolic: the representation based on convention only, no natural link between sign and object

Moreover, we consider this distinction as heuristic classification, where the icon and the symbol represent extreme values (see Fig. 11), when normally a sign has both qualities to some degree: the iconic quality to have semantically and/or signally proximity to the referenced object, as well as the symbolic quality, to draw to the object just by convention and repeated experience. However, we consider that such smooth transitions among the categories will be unproblematic in practice as the primary goal is not to uncover the essence of a sign but sensitize designers about the various opportunities to encode information by a sonic overlay. As follows, we want to discuss the three categories in more detail.

**Iconic mapping** An icon is a visible, audible, or otherwise perceptible representation of the thing for which it stands. In the auditory world, iconic auditory signs will be sounds that sound similar to the object [237]. Thus, the iconic character results from an imitation of sounds typically produced by the referenced object. For instance, the dog iconically barking refers to the barking dog, or the engine noise serves as iconic auditory of a moving car. Iconic sound design is typically used in a radio play, movies, and computer games to enrich the user experience. In some cases, weather owns strong iconic representation, like, for example, thunder. We aim further to uncover which iconic sounds and combinations of those are useful to incorporate in sonic overlays.

**Associative mapping** Going one step further beyond iconic representations, we can uncover associations that are reduced and linked to a distinct characteristic or feature. In the case of Starwars, Ben Burtt looked for familiar animal or machine sounds to establish credibility to ensure recognizable semantics for the sound effects: "The basic thing I do in all of these films [Star Wars and its sequels] is to create something that sounds believable to everyone, because it's composed of familiar things that you can't quite recognize immediately." – Ben Burtt quoted by Whittington [352].

Arbitrariness is based on some similarities between the sound and the referent but not as strong as in auditory icons at the iconic level. As Gaver [107] argues that in general, iconic/nomic mappings are more powerful than symbolic and metaphoric/associative mappings, because iconic/nomic mappings show a direct relationship between the auditory icon and the physical source of the referent.

**Symbolic mapping** "Auditory icons require there to be an existing relationship between the sound and its meaning, something that may not always exist" [205]. For example, this is the case if weather conditions do not come with literal sounds. A speaking example is the difference between thunderstorms and cloudy weather conditions. Whereas thunder offers an iconic mapping through its distinctive sound of rolling thunder, cloudy weather does not have such an explicit feature. In the absence of an iconic mapping, we ought to apply symbolic mapping, which "is essentially arbitrary, depending on the convention for its meaning" [107]. For example, when the VA announces cloudy weather, the consistent use of a particular sound establishes a symbolic relationship, similar to a ringtone that a user associates – over time – with a particular application.

## 9.4 User survey design and procedure

The first step in our design of sonic overlays is to define a conceptual mapping that is understandable by the users. Sonic overlays are more recognizable if they are based on iconic and associative mapping, with an active and purposeful linking between what is said and heard. This has the advantage, that no social conventions have to be previously established. Therefore, we conducted an online survey to collect associations with basic weather report events, such as rain (1), fog (2), frost (3), cloud (4), snow (5), thunder (6), and sun (7). In total, we received 33 complete answers but decided to incorporate also the described associations of 15 incomplete answers. We therefore collected a data set of 48 participants aged between 23 and 66 (male: 12, female: 19, non-binary: 1; mean age: 36,9 years).

Our survey did not aim at being statistically valid since it intended to sensitize our design phase. The survey was distributed in the area of Germany and Great Britain using social media services. All questions were not mandatory and open. We asked two questions for each of the seven weather conditions. First, the participants should name three concepts or terms they spontaneously associate with the mentioned weather conditions. Second, they should name or describe three associations of sound, noises, and/or music. Even if these associations are not explicitly set to music, they give the sound designers an impression of the semantic field that is evoked by each weather condition. Finally, we collected demographic information such as age, gender, and education. Afterwards, we descriptively summarized the results (see Table 12). Therefore, we clustered identical and very similar meanings. The table below shows the 10 most named concepts.

**Table 12**

Semantic and sonic associations regarding weather conditions. The numbers in brackets refer to the number of participants mentioning the respective association.

| Icon | Description | Conceptual Associations | Sonic Associations |
|---|---|---|---|
| ☼ | Sunny | warmth/heat (23), happy (10), brightness (9), sea/beach (8), ice cream (5), sweating (4), light (3), summer (3), holiday (2), cheerful (2), blue (2), sky (2), yellow (2), | birds chirping (27), songs (14), music types/styles (12), beach sounds/sea sounds (11), sounds of water/splashing (9), children playing (9), laughing/cheering (8), crickets chirping (5), individual instruments (4), high pitched noises (3) |
| ☁ | Cloudy | gray (12), dull (8), dark (6), shadow (4), sad (3), uncomfortable (3), probability of rain or snow (3), sluggish (2), windy (2), overcast (2) | music types/styles (20), sound of wind (9), silence (7), thunder (6), light wind noise (3), light water noise (3), traffic (3), string instruments (2), trees that blow in the wind (2), rumble (2) |
| 🌫 | Foggy | mysterious (4), mist (4), damp (4), headlights (3), cold (3), white (2), quiet (2), pea souper (2), epic scenery (2), darkness (2) | silence (12), a dark sounding horn (11), songs (5), muffled sounds (4), slow traffic (3), scary music (3), music types/styles (2), birds chirping (1), crow (1), echo (1) |
| ⛈ | Thunder | lightning (9), scary (5), waves/water (4), rain (4), strong wind (4), excitement (4), danger (4), bending/falling trees (3), thunder (3), dramatic (3) | howling/hissing/swishing (9), the sound of thunder (9), whistling (5), rain falling (5), drums beating (5), bangs (5), full orchestra (4), strong, whipping wind (3), sounds/instruments (3), rattling (2) |
| 🌧 | Rainy | wetness (20), puddles (6), raindrops (6), water (5), umbrella (4), cold (4), chilling (3), rushing (3), damp (3), pouring (2) | dripping (28), splashing (11), water rushing (6), footsteps in puddles (5), songs (5), pattering (3), drumming (3), music types/styles (3), opening umbrella (2), cars going past on wet roads (2) |
| ❄ | Freezing | ice cracking (18), coldness (13), white frost (6), freezing sounds (6), ice (5), slippery (5), danger (5), dressing up warm (4), clanking (3), single instruments (4), snow (2), | music types/styles (8), crunching (5), ice skates on ice (5), shivering (4), scratching (4), scraping on cars (4), slipping and falling (2), songs (2), |
| 🌨 | Snowing | white (11), cold (10), snowman (7), brightness (5), silence (5), calm (4), winter (3), snowballs (3), winter sports (2), flakes (2) | (snow) crunching (16), Christmas/winter songs (11), silence (12), ice skates sliding on ice (3), Christmas music (2), clumping (2), shoveling snow (2), crackling fireplace (2), soft music (2), muffled noises (2), bells (2) |

## 9.5 Results of semantic and sonic associations

### 9.5.1 Iconic mapping

The survey showed that for the specific weather events, participants had varying difficulty associating tones, sounds, or music, and these associations could be a lot diverse. The association turns out to be most coherent where there is a natural iconic mapping, i.e., where a weather event naturally causes sounds. Rain or flashes represent a fitting example, therefore. In the case of rain, for instance, the associated semantic field revolves around the theme of wetness, water, and raindrops. Those are also associated with certain moods such as chilling, and uncomfort but also calmness, and certain colors such as dark and gray.

The theme of rain and raindrops can also be found in associations such as pouring, as well as in associated objects for personal protection, such as an umbrella. The sonic field translates the theme of the semantic field of water in terms of nature sounds caused by rain, e.g., splashing, water rushing, dripping, pattering, and drumming. Besides, mainly naturalistic or nature-simulating associations were named, e.g., thunder and lightning, running faucet, or rice grains weighing back and forth. In the case of lightning, an iconic mapping is found in most cases that the electrical discharge produces not only lightning but also thunder. This fact has led to a quite homogenous sonic field, where most participants directly associate thunder or specific forms of thunder such as dumpling, crashing, or banging. Furthermore, it turns out that lightning is semantically and sonically associated with rain, expressed, for example, by sonic associations such as drumming, waterfall sound, or pattering rain.

### 9.5.2 Symbolic mapping

The opposite case presents weather events where a natural iconic mapping does not exist. The most prominent example of such a case is the cloudy weather. In contrast to rain or flashes, the participants do not associate specific natural events or activities but a vague, general impression of gray, dark shadows, coldness, and a quite unspecific, melancholic mood of discomfort, sluggishness, and bad temper. The theme of coldness was also expressed by mentioned protection means like bringing a jacket or sweater weather. Occasionally there are associations with seasons, e.g., autumn or places like Germany, as well as activities such as doing sports outside or city trips. This broad, unspecific, and, as it were, the soundless semantic field is echoing in the sonic field. Here we observe the heterogeneous answers that aim to differently translate the vague ideas of gray, gloomy, and melancholy sonorously. In addition, two participants answered the question about associated concepts but omitted the question about sonic equivalents. The other answers show a wide range of sonic associations. What is striking here, is the frequent tonal characterization of cloudy by general musical characteristics (melancholic music, ponderous beat, polyphonic male choir), certain musical trends (lo-fi beats, jazz music), or individual instruments such as strings, and styles of sounds such as muffled sounds or dull hum without associating one specific sound or piece of music.

Participants mentioned natural sounds such as wind or water less frequently. We also find it inspiring that some participants associated human noises like sighing, breathing, and the sound of yawning to give the melancholic mood a sound.

### 9.5.3 Inbetween iconic and symbolic

The answers further show that most associations cannot be unambiguously classified as iconic or symbolic mapping, but mostly represent something in between. Therefore, in our view, it makes sense to understand the schema outlined in Section 9.3.1.2 as a heuristic rather than a strict category system. Sunny weather is one of the examples where iconic, associative, and symbolic mapping is balanced. In the semantic field, we see strongly iconical responses, e.g., warmth/heat, brightness, and blue sky, but various answers more indirectly related to sunny weather such as summer, expressions of summer feeling like being motivated, happiness — for instance, expressed by laughing – as well as diverse summer activities such as cycling or eating ice cream. In addition, some answers refer to measures for sun protection, e.g., sunshades or sunscreen. The corresponding sonic field also reflects this semantic field. Unlike flashes or rain, the sun does not directly cause sounds. The associated natural sounds are not iconic but rather associative. Various participants mention sonic expressions typical for a sunny sea holiday, such as the sound of waves, splashing in the water, or voices at the beach. These associations present indexical signs in the sense of Peirce [244] because of the causal chain of the sun (causes hot causes refreshing beach holiday causes sea sounds).

By the same token, they present a metaphorical mapping in the sense of Gaver [107] because in western societies beach sounds become a metaphor for a hot summer, good feeling, and sunny weather. In addition, some participants associate sunny weather with crickets chirping or birds chirping. Again, there is an element of indexicality and metaphoreness in these associations (as not sunny, rainy weather physically impedes both, chirping and singing, and so both natural events have become metaphors for a sunny summer). Less indexically but more metaphorically are answers such as laughing or cheering. While both are not directly metaphors for sunny weather, they are metaphors for happiness and good feeling — which was one of the associations in the semantic field. This feeling of lightness, sunny weather lifestyle, and good mood are represented by many musical associations, both regarding styles (light pop music, light electronic music, major sounds, reggae, Latin American music, as well as regarding particular songs such as “Sunshine Reggae” from Laid back, “O.P.P.” from Naughty by nature, and two ice cream commercials, “So schmeckt der Sommer” (Engl. “This is how summer tastes”) and “Like ice in the sunshine”.

Overall, the answers indicate that iconic mapping is prominent when the weather event causes typical, easy-to-remember sounds. In contrast, when those sounds were not available, participants suggested symbolic mapping more often.

## 9.6 Developing a library for sonic overlays

Our library for sonic overlays is based on the empirical and descriptive results of the survey described in Section 9.4. Further, we use the categories of iconic, associative, and abstract sound to cluster the results and produce sound clips that show a high discrimination quality for all seven weather types. We will explain our design rationale and according steps as follows.

### 9.6.1 Design Approach to Enrich Alexa's Weather Report

In contrast to Mynatt et al. [224], we decided to gather conceptual mapping and physical parameters by a free-form survey before the design phase. Further, our goal is not to design auditory icons but to illustrate speech by using iconic, associative, and abstract soundscapes that are not synthesized into an identifiable sound-only design but serve the purpose to illustrate spoken information.

The seven most distinctable weather types were chosen to be the core of this design: sunny, rainy, cloudy, foggy, snow, frost, and thunderstorms. The authors sorted the responses into categories depending on each sound's connection with the weather in question:

- Iconic sounds, which are caused directly by the weather
- Associated sounds, which are expected to occur in conjunction with the weather but are not directly caused by it
- Abstract sounds, which have a connection to the stated weather type in the respondent's mind but are not necessarily linked to it

This categorization is based on previous conceptual considerations, as introduced and explained in Section 9.4, and enables easier identification of any positive or negative reactions to certain types of sound by users. Further, we want to highlight that, in contrast to rain, certain weather conditions like foggy and cloudy have no iconic sounds. This must be taken into account when creating the respective soundscapes. It will also provide an opportunity to evaluate how a lack of iconic sounds affects the user's overall perception of the soundscape.

Therefore, as a first step, we categorized the survey results described in Section 9.4, sorted from most common to least common, for all seven weather types (see Table 12). Table 13 exemplifies a rainy weather condition (see below).

### 9.6.2 Structure and elements of sonic overlays

Sonic overlays and earcons/auditory icons share multiple features, such as conceptual mapping and encoding information by sounds. Yet, the survey of Cabral and Remijn [40] shows that in contrast to sonic overlays, earcons are quite short (mostly between 0.5 and 3 s). As

**Table 13**
Sorting and categorization of survey results using the example for rain.

| Iconic | Association | Symbolic |
|---|---|---|
| Rain noise | Footsteps in puddles | Car horns |
| Dripping | Tyres on wet road | Many voices in closed space |
| Splashing | Opening umbrella | Drumming |
| Rain on the roof | Rustling raincoats | Boiling water |
| Storm noise | Wind blowing | Tapping |
| NA | Drains gurgling | White noise |

our sonic overlays attempt to illustrate speech-based information of VAs, we need to take into account that talking often lasts from a few seconds to minutes. For instance, the weather report of the German Google Assist takes about 10 s, allowing sound designers further options regarding rhythm, using pauses, proving ambient sonic overlays, and other temporal parameters. Another main difference is that in sound overlays, the voice conveys the primary information, which liberates sound designers to more subtly encode the information and, for example, emphasize or ironically comment on the spoken information by sound. However, it also creates new constraints, such as that the sound overlay should not interfere with the voice making it difficult for the user to understand what the assistant has said.

The examples created were each around 25 s in length and incorporated sounds based on the most frequent answers given in the survey, in combination with a synthesized voice similar to that which would be heard from a VA. Further, a proper difference in loudness between the soundscape and speech ensures the discrimination quality within the sonic overlays. The structure of each sound overlay clip was consistent across all the weather types: each starts with around 5 s of sound effects to build up a soundscape representing the weather, then a voice would explain the weather condition and temperature, followed by additional 10-15 s of audio. If the clip includes any musical elements, these are incorporated into the soundscape after the voice has spoken.

Musical elements and soundscapes are essential to creating an expressiveness of information that speech could not. Two examples also incorporated musical elements, besides sounds and spoken words. The example for the sunny weather condition incorporated a guitar melody inspired by “Here Comes The Sun” by The Beatles, as this song was mentioned by multiple survey respondents in association with sunny weather. Further, the example of the frosty weather condition incorporated an original melody using tones and timbres identified in the surveys as conveying a feeling of cold, icy weather. When creating the soundscapes, sounds with a rather direct connection to the weather type in question were prioritized, e.g., the sound of wind or falling rain. However, in impossible cases, more abstract sounds were preferred instead, e.g., the cloudy soundscape that featured heavy traffic noise. In either case, all sounds featured in the soundscapes were selected from the survey responses.

## 9.7 Evaluation of the sonic library

### 9.7.1 Interview Study Design and Procedure

Frequently, associations and imagination are linked to prior experiences and their cultural background [152]. Therefore, we did not aim at a statistical representation of the populations in Germany and Great Britain. We were looking for participants with heterogenous cultural backgrounds able to speak and understand the English language. For recruitment, we used snowball sampling in our extended networks [23], thus, we posted requests in social networks like Facebook, international telegram groups and private messenger services. To further diversify our sample, we asked the first participants for references from their extended networks. Most of the 15 participants (4 male, 11 female), currently lived either in Germany or the United Kingdom, in addition to one participant living in France and one in Palestine. However, their geographical backgrounds were significantly more diverse, including south-east Asia, Sri Lanka, Canada, and Russia, among other countries. This diversity in backgrounds helps identifying how a person's current or past environment might affect the evaluation of sonic experiences and weather types. Table 14 provides an overview of the corresponding data regarding age, gender, and current and previous residence. Most interview participants had at least some previous experience with VAs . Participants who were inexperienced in interacting with VAs had a basic understanding of how they work. Therefore, we only explained the sound overlay concept. Participation in our study was voluntary and did not involve any compensation.

**Table 14**
Study participants (n=15) representing international differences in culture and residence.

| ID | Age | Current residence | Additional info | Previous residence |
|---|---|---|---|---|
| P1 | 27 | Germany | Industrial, edge of forest | Hong Kong |
| P2 | 29 | UK | South England, rural | / |
| P3 | 23 | UK | Hull, suburb | Used to live in a more rural area |
| P4 | 62 | UK | Scotland, coastal | Used to live in New Forest |
| P5 | 57 | UK | South England, countryside | / |
| P6 | 60 | UK | South England, countryside | / |
| P7 | 28 | Germany | Industrial, edge of forest | Azerbaijan |
| P8 | 25 | Germany | Industrial, edge of forest | Toronto, Canada - less rain, colder in winter |
| P9 | 67 | UK | Guildford, leafy suburb | Sri Lanka |
| P10 | 20 | France | Small town, near coast | / |
| P11 | 23 | Germany | Rural, edge of forest and small town | Spent 3 months in Canada |
| P12 | 25 | Palestine | Varied seasons, hot in summer, rainy winter | SE Asia - weather very similar all year |
| P13 | 46 | Germany | Small mountain town | Village in Lower Saxony |
| P14 | 33 | Germany | industrial area, city | St. Petersburg, Russia |
| P15 | 31 | Germany | Industrial, edge of forest | / |

We chose a qualitative interview study approach to explore the subjective perception and usefulness of the sound overlay library. Each participant listened to both conditions: VAs with speech only and VAs featuring speech with sound overlay for three randomly chosen weather types. We created a randomized experimental design without repetition, so that each participant was played two of the three sounds, e.g., weather report with/without sound overlay for rain (1), fog (2), frost (3), cloud (4), snow (5), thunder (6) and sun (7). First, randomization without repetition ensured that at least six subjects listened to each of the seven weather reports. Second, the randomization was intended to minimize a sequence or order effect. The experimental design randomized the order and also the combination of the other samples (e.g., with snow and storm or with frost and sun) to account for possible changes in opinion brought about by hearing particular examples in combination. Additionally, the order of the clips for each weather was also randomly selected, taking into account that listening to the first clip might influence the next. We uploaded the sound library to youtube to share only the chosen links to the clips during the interview. After listening to each clip, the interviewee was asked specific questions about what they had just listened to, followed by more general questions about the concept and their impressions of it, e.g., did you recognize the sound as the correct type of weather? How long did it take? Or did the information come across, and how does it make you feel? Each interview lasted around 35 min on average and was conducted over Zoom.

Finally, the interviews were transcribed verbatim and coded inductively and independently in MaxQDA by two researchers using thematic analysis [122]. We focused on the effective sonic experience of the weather types and the perceived differences in design and usefulness. Also, we explored the impact of combining speech and sound and its implications for structuring and contextualizing information.

### 9.7.2 Findings

Some participants regularly used VAs to check weather forecasts but the majority relied on websites or smartphone apps instead, usually citing the level of detail offered as the reason why. Several stated that the short spoken summaries by VAs did not give enough specific detail to plan a whole day.

**9.7.2.1 Supporting imagination and experience** Sonification aimed to support people to produce images in their minds that use emotions and prior experiences associated with distinct and ambient noises. By using the examples of weather, we could observe clear challenges in design for two specific groups of weather types: almost silent events like fog, sun, frost, and cloud, and loud events like rain, snow, and thunder. Although the prestudy foreshadowed possible challenges to design recognizable and unambiguous soundscapes, the cloudy weather seemed to cause the majority of problems in correctly understanding the presented information.

Most of the participants responded to the idea positively when listening to the samples and expressed vivid accounts of their imagination. Some welcomed their emotional responses and explained that this makes the interaction less boring and monotonous but more dynamic

(P1). This evokes a space “like being on a boat in the ocean” (P3), when listening to the audio clip of ’fog’. According to P1, weather reports supported by soundscapes felt less “artificial” than speech-only and created a kind of “haptic feedback” of the information:

> “I think it’s more emotional because you do have like, an image, sort of, in your mind. Yeah, I like the fact that it’s not only rain. It feels like car and rain or some background noise. You know, it feels like you are really in the middle of the city. And you don’t have an umbrella, and you are suffering from a pool. (...) In this context, I think you want to use a temporary, really precise message of the weather, and I think this achieved their goals.” – P1

In particular, the soundscapes emphasized typical feelings associated with specific weather conditions, as participants explained that the thunder sounds made them anxious (P3), a sunny city equaled good feelings (P2, P7), or freezing temperatures indicated not to go outside:

> “So we were like heavy winds, which were full of crystallized snow. And you could hear yourself like walking through the huts. Cold, like the freezing or the snow, which feels like the ground. And, yeah, the wind was so strong that you did not want to go outside at all.” – P14

The soundscapes of pleasant and unpleasant imagined situations alike enhanced the intended message and supported possible adaptations of the participants’ behavior, like being motivated to go out (P8). Some saw the concept particularly useful for special occasions and ambient background information needs (P13). Moreover, P1 and P14 reported that the sonic overlays contributed to a calm and relaxed feeling.

> “Natural sounds in general. Also the crows and animals and things like that. Because sometimes people are stressed about everyday life or life pretty often. So they have, they want, like something to relax. And maybe one selling point of this app or a voice assistant would be like that one can relax, that are in our everyday life.” – P14

Sound is not considered overall necessary for a system solely designed to give factual information (P12). While regular forecasts are unbiased, sound adds a character to it that can have positive or negative connotations. This can help to form decisions based on the weather because it is easier to imagine yourself in the context. P12 indicated that the specific information might not be as memorable, but the overall impression was much stronger and helped with understanding the consequences of the weather conditions. Another piece of feedback from several participants was that the soundscapes made it easier for them to visualize the weather and think of how to prepare for or react to it. P3, P11, and P10 considered this useful for morning routines or directly after waking up in a dark room. Moreover, P10 calls the design concept more reassuring by giving a feeling of naturalness and coziness (P10). P3 also was surprised that it was not already commonplace for VAs since visual apps use graphics to add more context and to communicate information in a more appealing fashion (P6).

Further, this concept bears a chance to give friends coming to visit a more precise idea of the weather conditions and makes it more interesting to share (P13). Additionally, it might help to feel a deeper connection and experience with the represented location if you live far away, as long as the information represents the reality:

> "Let's say I want to go to London and I'm checking the weather in London. Or maybe I want to see the weather in a different country right now. For a particular reason, it is important to me. (...) but instead of saying rain and the strength of the rain, it might add more because if it is on real-time as opposed to a forecast, if it is music, then I feel it. This level of, you know, the burden of interpretation. But if they are actual, it's almost as if they are giving real-time Information. Then if they are making me hear it, how it is, how snow is flowing. They know how it is raining in London or wherever away from the I can see from my window. I can see data that has been an interesting dimension that I would be interested to see." – P9

Meanwhile, missing experiences of weather conditions or landscapes might contribute to misleading interpretations or less precise perceived information. For example, P15 could not recognize and relate well to the foghorn sound that represented foggy weather in comparison to P4, who imagined their current residence:

> "I could picture the coast where I live, which is a harbor, small harbor and the sea and foggy sea and the fog coming into onto the land, which it does where I live (...) quite often. So yeah, a totally foggy, virtually visible. With the emphasis on sounds that you hear rather than what you see." – P4

As P1 grew up in a large city, hearing footsteps in the snow made it difficult to differentiate between snowy and frosty weather and carried over all the impression of a hiking vacation in nature rather than an intuitive sense of the weather conditions. She was missing the noises of traffic, for example, cars. In contrast, P6 noted not to include traffic noises because those do not symbolize sunny weather to her. In a similar vein, P8 and P13 did not consider children playing outside as an appropriate illustration of sunny weather, and hearing splashes reminded P13 rather of rain.

**9.7.2.2 Sonic information design** The sonification of information relies on abstract and iconic sounds, as well as relevant music pieces and speech. Particularly abstract sounds contributed to an active imagination and conveyed the meaning of the weather conditions. Therefore, all participants pointed out that the incorporation of related sounds gave a better impression of the scenario:

> “I think all of these have given me very if you’ll pardon my illusion, Animal Crossing kind of vibes. I don’t know if that was a deliberate image or just circumstantial. But it’s not the weather. The tones fit the weather, the sounds of the light. With this one, you could hear like it was like birds singing. Nice day, kids having fun. Like, I think that was a roller coaster. And then the marimba at the end or like a guitar.” – P12

Overall, the concept does not represent a simple sequence of symbolic sounds. Hence, the soundscape has to be layered with consideration. An urban environment might sound different than pure nature but it has an equivalent impact when sounds like background noises are combined that indicate events happening during this kind of weather or the place of experience.

> “I like that. Not just the sound of it. It really sounds like you try to mix it with different elements like the surroundings. Sometimes the sound is not really directly about the weather, distinctive. But I think that’s really awesome. Some feedback is that, for example, there’s the second one I have the most problem understanding. The foggy one.” – P1

The participants appreciated incorporating musical elements that acted in a similar vein to convey information and emotions that noises could not. For example, P11 stated that music represented “icy” conditions much better than footsteps. Likewise, this type of sonification supports the differentiation of similar states like frost, ice, and snow. P2 explained that music was thematic and indicated light and pleasant snow by that:

> “I think it was very thematic in the sense that it gave you an idea of what to expect. It kind of indicated it’s going to be like, you know, sort of like, oh, it’s nice. You can walk in it. It’s going to be like pleasant thunder. It didn’t seem to be indicating snowstorm: Stay in your house!” – P12

Likewise, the use of a guitar, for example, may produce a “calming effect” (P8). In contrast, P11 described VAs as a convenience and aimed for efficient interaction, where music might be in the way. Further, P12 was concerned that not everybody would appreciate such a design decision as well:

> “I liked it. I mean, again, it’s I think the sort of people that would be put off by the extra fluff at the end. People that would just look at a website and wouldn’t use the service anyway. So I think it’s adding an additional level of sort of engagement to people that are going to be using the product.” – P12

However, the music proved to be an effective element for supporting imagination and speech-based information:

> "All the right information came across straight away. And what was interesting was that because I'd heard the music first, I had this same image of this road going into the distance and everything, a little bit orange. Don't ask me why, but maybe going into the sunrise, sunset, you know, a pleasant travel image, basically." – P4

An overall trend in the results was that soundscapes that more heavily used iconic sounds were more well-received than those which relied solely on abstract sounds. This presents an issue for weather types that do not have any associated iconic sounds, such as cloudy or foggy. Especially iconic sounds are well suited to represent precise information, entail clear messages, and evoke past experiences as associations at the same time. Further, natural sounds are closely tied to the expectations of weather conditions:

> "And because of the sound of the birds, you kind of feel it's sunny and the kind of feel that people outside and that things are happening outside. So you assumed your kind of mental image was this sort of like sunnier, drier weather." – P9

In comparison, particularly rain and thunder were tangible noises with high and quick recognizability. Participants (P13, P3, P2) discussed afterward, for example in the case of snow and frost, how the granularity of weather conditions and their differentiation could be supported by a variety of iconic noises.

> "And as I mentioned before, you could play a different thing. So the severity of it. So you've now winden and instead of sort of a lighter sound, but more heavy, I assume they were sort of sleigh bells or reindeer to indicate a more hazardous conditions maybe. Yeah, but yeah, I know it was all very easy to hear that it gave across everything you were trying to say." – P2

P13 added that it could be confusing if there are snow sounds but only 50% chance of snow, for example, and that it may be better to build up from a wider bank of sounds for variations (P3), for example, a concise representation of temperature and that "Rain sounded maybe not as 'heavy' as the voice said" (P3).

Besides, difficulties arise with sounds that cannot be represented iconically because of the absence of noises, for example, with sun, clouds, or fog. However, this might lead to confusion by trying to substitute by using crows or horns that occur or are used in cases like fog. P11, P10, P4 and P1 had trouble understanding the meaning of crow noises and considered them as confusing.

> "Then I think they were crows or rocks, the birds. For me, they could make that noise. Morning. Evening. Any weather? Probably. But then not everyone's going to know that I live on the coast. And for me, I was wanting to. Seagulls, of

> course. But of course, it's not everybody lives on the coast. So, yeah, it wasn't a big deal, but the phone comes with a real positive clue, so it didn't matter. The rest was just atmospheric. Quite nice to listen to." – P11

At times, some participants (P8, P1, P9) felt overwhelmed by the combination of too many sounds and suggested cutting back (P8). Musical elements could of course be added but also detracted from the message and would leave just a noisy impression (P9). Overall, the balance of iconic and abstract sounds provided an enhanced experience and emphasized the information. Nevertheless, the design should focus on communicating a clear message as well:

> "I liked that they didn't all do the same thing. So you had some that were the literal sound of the weather and some that with sounds associated with the weather. I liked that there was a bit of a mix. I didn't like that, I didn't feel like any of them gave a clear communication of temperature. (...) I liked the sounds there and I liked the length of them." – P3

**9.7.2.3 Adding Sound to Speech** Besides iconic sounds, a deliberate choice for the design of sound overlays was to incorporate speech providing precise weather information. Many participants claimed that without speech, they could not identify the correct weather conditions, especially concerning fog, frost, and clouds:

> "Well, what I noticed is that the abstract sound only came after she talked. The voice (...), there was no ambiguity. And I really knew that it was the frost that made the sound." – P11

In contrast, some participants indicated that in the case of rain, the speech felt even unnecessary, and, in the case of thunder, it was even more clear than vocal information:

> "I felt like it basically brought things across. The voice said heavy thunderstorms. And I feel like maybe the rain wasn't heavy, heavy, heavy. But at the same time, that would raise the question of, well, how many different words does a voice assistant use when describing weather? And then can you map all of those words on to a sound of rain, like the thunder sounded heavy?" – P3

Overall, the intended and sonificated meaning of rain, sun, snow, and thunder was recognized most frequently and almost immediately. P11 added that by the sound he imagined, it is even easier to remember to bring an umbrella. Further, P12 explained by listening to thunder that he had clear thoughts on the preparation for the upcoming stormy weather.

> "I think it was like supporting the voice. Sometimes I also think that the voice was completely unnecessary. In extreme beavers and extreme weather conditions, for example, when it was like snowing or raining. But a service (...) it will

> be like necessary to at least say the temperature. And I mean the information about that it's snowing." – P14

However, most participants considered speech for quick and precise information, like temperature indications (P14), valuable, especially those participants who might be impatient because they are in a hurry (P14, P1, P9). Furthermore, participants feared that voice and soundscapes could compete for attention sometimes, e.g., because of false expectations regarding the timed structure:

> "Since, I think, it's one minute. Whenever, (...) it's not necessary, but it can be of it can be a bit frustrating if you missed the moment that it starts saying." – P10

Further, P10, P13, and P12 expressed concerns that voice and background noises were overlapping too much, e.g., children screaming while playing outside (P13). Hence, despite a better image of a complete scenery, speech-based information was drowning down:

> "In the same instance, you get like in films, sometimes there's a dialog scene. And then the orchestral score or the things in the background is so loud, you actually can't hear what's going on, which then detracts from the product, which I think is something you guys have managed to avoid." – P12

Additionally, P11 mentioned that sound shouldn't seem to contradict speech to not add to ambiguity and confusion:

> "It doesn't add more information to this, to the stuff that she's saying. Because in the first part of the snow, it added snow. She didn't say anything about snow. And the second one added wind, even though the voice just said it's foggy, not windy. And it must be very difficult to achieve. But I think that's really important that the sound is very much in line with the words and not adding or taking away information." – P11

P11, P8, and P9 stated that the use of sound elongates the application and requires patience. Consequently, in their need for quick information, they would prefer speech-based, either through voice or by glancing at their phones.

> "In a car. Probably like when you need to just have the information (...). But when my mind is like, I just want to know this and then I want to do something else. I don't know in which situation that's the case. Usually, most of the time, but when I ask: 'Okay, what's the weather going to be like?' And then they tell me and then I cannot ask another question for like 5 s because I have to wait until the rain stops. That would annoy me so much." – P11

In total, we could observe balanced opinions on the preference of voice or sound–first regarding the structure of the sonic overlays. Therefore, some of the participants (P6, P14, P9, P11, P5) argued, for example, P14 and suggested starting with speech first when designing sonic overlays:

> "I think it will be better to start with a voice or maybe a millisecond off or a nanosecond. I'm not sure of like of forever, of a silence and then the voice. Because I think sometimes people don't have patience. Some people don't have the patience for waiting until the voice pops up." – P14

P6 demanded to have speech instantly - "facts not thrills" - but could imagine maybe a short sonic fade-in before and fade-out quickly afterward. A further advantage of speech–first might be reduced ambiguity and sound as additional layers that can be better interpreted (P9). P11 suggested making the clip shorter overall to make it more efficient, although this might lead to impressions that interfere with the voice.

> "Waiting in suspense for the voice - then it happens suddenly. Voice and sound should start at the same time then let the sound carry on for just a few seconds afterward to leave an extra impression." – P11

On the other hand, participants had found reasons to start with sound as well:

> "No, I think the fact, that the lead-in was an audio clip of the weather type or something alluding to the weather type followed by the information, then followed by another weather clip with a bit more music. I think it gave you an idea of what was coming. It was then clarified and then you got this sort of little ribbon on the top of whatever you're referring to us." – P2

Many participants appreciated the current design structure of the sound overlays. They pointed out that sound introduces impressions and scenes as afterward speech fades in to confirm and clarify weather conditions. Besides, P10 describes this design as feeling less aggressive than the assistant speaking at you immediately. Nonetheless, participants like P14 and P4 emphasized that this concept needs time to get used to it first.

**9.7.2.4 Sonic Contextualization of Experiences** The sonification of information might be extended to other applications and design spaces, as the statements of our participants show in the following. However, they expected some limits regarding the usefulness and experiential value. In particular, situations that allow for ambient sound and personal moods that welcome entertainment, e.g., driving in the car or waiting in general. For instance, P8 considers background ambiance, like the sound of a fireplace or ASMR (Autonomous Sensory Meridian Response) for cooking or studying as relevant. P13 would consider hearing the sounds of frying/chopping, etc., to be more amusing. Additionally, P4 describes a possible situation at work:

> "When I'm working in home office, I'm able to choose. When I go out for a walk, I could look out the window. But in Scotland, that won't tell you. You really need to know that temperature, preferably what it feels like. I mean, that's peculiar to Scotland. It doesn't set up. The temperature is what you really want. And yes, I could come out to whatever I'm writing or reading. And I could click or met Office, and I could get it. But if I could just get it instantly, you know, like that just: 'Oh, I wonder if I need a hat and a scarf as well as a coat today. Do I need two pairs of gloves or one?' Then I would quite like that. A fun way of doing it, especially as I want to then forget about work, although I actually associate my laptop with work. So for me, just to have some quick little sound, and off I go for my walk" – P4

Besides asking for the weather or specific information, the news is a frequently used service of VAs and radios alike. However, our participants had contradictive thoughts on the sonification of this offer. P4 could imagine a benefit of applying sounds to the presentation of traffic updates, travel reports, or election/sports results, especially at times you want to know the info in a flash. In contrast, P1 expressed that sound might distract or manipulate information. Further, for P3 bad or scary effects might be reinforced.

> "Honestly, I, I don't, I cannot think of anything that would benefit from that. Because it always conveys some sort of interpretation or maybe opinion or emotion. So if you add it to a news article, it's not neutral anymore. And I read the news to make my own opinion. So I wouldn't like to be presented with somebody else's emotions." – P11

Whereas more participants can see potential design spaces at home by enhancing other media and smart home applications. P3, for example, would wish for audible feedback on loading times and completion of tasks. P1 explains in further detail how a sound or earcon library of a current VA might enhance the notification experience of deliveries:

> "Alexa might have some sound ding ding on this topic. Another possibility is when I'm anticipating a package, I know the different stages of the package, like, is it a ship that is delivering (...). It will be quite helpful because right now, they treat it as a notification. Like maybe you have, you can extend these to some parts of: 'Are going to arrive today'. If they can have a different sound to describe where exactly my package is." – P1

## 9.8 Discussion and implications

In light of our research questions, we want to discuss our results and provide implications for the design of future voice interaction. So far, Alexa is seen as Voice Assistant, very neutral in their answers with little capabilities to express emotions [53]. A significant amount of research in the fields of speech science aims to address this shortcoming, respectively emotional speech and voice design [198, 347, 1, 169, 288, 286, 187, 37, 87]. In this paper, we complement this area of research by outlining a supplementary approach, using sound as a modality that could add a new dimension to voice interaction and enrich the user experience. In particular, we focus on the relation between speech and sound and the balance between communicating information and inducing emotions through sonification.

### 9.8.1 Sonic encoding for voice interaction design

**9.8.1.1 Building soundscapes** The prevalent design paradigm regarding sound is to precisely encode information to substitute functions and representations [27], leading to different kinds of auditory icons and earcons that are highly recognizable. However, that also requires either a clear sonic representation, or users to learn its meaning first. As with current VAs and computer systems in general, we can observe the use and purpose of earcons to signal warnings or direct attention to events on short notice [27, 107]. However, iconic sonification might come at the expense of rich soundscapes capable to transport emotions, atmospheres, and further experiential qualities, as known from the design of classical media and extended realities [152, 271, 46, 150].

Extending the purpose of sound by substituting single functions and representations, our results indicate that sonic overlays may support voice interaction to encode, illustrate and communicate messages. The combination of iconic, abstract, and symbolic sounds shows a positive impact on the perception of weather reports by speech-based interaction. Participants described their experience as stimulating and entertaining, quite the opposite of previous experiences with VAs. Thereby, iconic elements support the recognizability of intended messages. Some weather types gained noticeably less positive feedback than others, particularly weather types that relied more heavily on abstract sounds such as cloud and fog. As these require the listener to draw connections between the sounds and the weather in a less direct way, they are more open to interpretation and have more potential to cause confusion. These potential issues first appeared as early as the pre-survey; these weather types had fewer associated sounds suggested overall, and the most common response for a sound associated with fog was “silence”. Musical elements as well as abstract soundscapes serve as an illustrative layer to build a holistic impression of the specific weather conditions and are a carrier for moods and emotions. However, a missing combination of iconic sounds might obscure some information.

With our work, we present a structured design approach to sonificate and illustrate voice interaction and, thus, enrich the experience of weather reports. So far, only a little work on methods and research regarding design approaches of voice interaction, especially in combination with sound design, exist [298, 49]. Current approaches to voice interaction design are based on collecting example dialogues, spoken terms, expressions, and paths as design materials. Similarly, we collected associative mappings for each message of a weather event and categorized those into abstract, iconic, and symbol design elements to develop a not exclusive sound library. Although the design was well appreciated, we need to balance abstract soundscapes that affect the experience with iconic sounds, meanwhile ensuring recognizability of the intended message to communicate information successfully.

**9.8.1.2 Layering sound and speech** As our results indicate, the sonification of interaction opens the design space for more ways of expression [298, 49, 271]. However, voice remains a precise channel to communicate information and is perceived as an efficient and convenient way of interaction. Therefore, participants expect sounds to illustrate exactly the information of the voice channel and avoid contradictions from both channels. Further, by using abstract concepts like "children playing outside in the sun", designers have to be careful not to mix channels in parallel that entail soundscapes based on human voices. Otherwise, the discrimination quality is not guaranteed. Besides, more research into differences in similar weather types like frost and snow could prevent misunderstandings. However, participants were skeptical whether, e.g., 50% probability of rain, could be communicated via sound. Yet, they still desired a high granularity to express the characteristics of weather conditions.

The structure of the audio clips regarding the temporal position of sound and voice received mixed feedback from the participants. Some liked the structure of starting with the sound, then introducing the voice, and ending with more sounds as it gave them time to form an impression of the weather from the sound that later was confirmed and clarified by the voice. However, other participants felt that the clips in their current form were too long and that they wasted too much time compared to a voice simply speaking the weather forecast in just a few seconds. Although almost all believed that the sounds produced a better connection to the weather than the voice alone, several interviewees indicated wanting to hear the voice-first to get the most information as quickly as possible. However, a more matching combination of both might reinforce the impression that the sound illustrates what the voice was saying in real-time. Currently, the voice simply speaks over the soundscape after a few seconds.

Overall, sonic overlays illustrated and strengthened the voice message. Speech added the preciseness of information, especially for events or impressions that naturally are silent and hard to sonificate. Besides, a certain granularity and discrimination quality in sound design might positively impact the preciseness of information. However, the temporal position of sound and speech has to be purposefully integrated into the overall design and needs more research to give clear implications.

### 9.8.2 Balancing emotion and information

**9.8.2.1 Authentic soundscapes** Data sonification may serve both purposes, conveying information and emotion [271]. Sound design in Science Fiction gives the future a voice, linking the effects to the imagery to enhance the credibility of the cinematic reality [352]. The same holds for the role of sound design in games and XR [150]. Oftentimes, the goal is to create new worlds and experiences that are not nonexistent or less prevalent in real life. This was quite the opposite for our study because participants expected to understand the sonic overlays effortlessly. The main goal shifted to imitate the surroundings of known places and build on past experiences to encode information. As our results indicate, social context and personal residence environment greatly impact the upcoming associations and respective interpretations. For example, people who live in big cities might practice hiking as a seldom leisure activity, whereas people from the countryside might have a distinguishable understanding. The same applies to cultural experiences, e.g., festivities like Christmas associated with specific music and instruments. However, besides supporting the imagination of the known, places in different parts of the world can be illustrated in the same way. Yet also, in this case, it might be perceived as more worthwhile to experience representations quite close to the original experiences of people living in those areas.

Finally, experiences could be even further personalized by using location data, information on the surroundings in this area or during the daytime, and other chronic data to match the experience of the area. This approach would allow for enhanced recognition of sonificated information and for users to empathize with new places and experiences.

**9.8.2.2 Encoding emotion** So far, VAs lack an engaging experience that motivates users to interact on a regular basis [53] and are regarded mostly in utilitarian ways by users. Following the call of researchers to explore potential experiential qualities of VAs [298, 49, 61], speech science research [198, 347, 1, 169, 288, 286, 187, 37, 87] aims at encoding emotional information and expressiveness into the sound of voice and the way of speaking. With our alternative design approach, we investigated the design space to develop and promote an expressive context for dubbing, voice-overs, and future voice acting [52, 364].

Furthermore, our study focuses on exploring the various options to design surrounding and ambient sound contributing to the affective experience of VAs. Our results indicate that sound overlays could enhance imagination in comparison to voice-only interface design. Moreover, our participants reported both calming and anxious effects that either feel relaxing, or symbolize and promote action. This is also due to sound building up a closer complete scene, making it easier to visualize and respond than simply hearing words.

In the tension field of expressive and informative interaction, designers act responsibly and consciously regarding the sonification of positive and negative experiences. As our data shows, some participants were concerned about manipulative misuse of sounds, for example, when discussing news as further context for sonification. Clearly, some prefer “facts not thrills" (P6) and want their information not emotionalized. Further, some users deliberately do not want that triggering of negative feelings. Therefore, designers might also aim to balance hazardous weather conditions like thunder with sounds that indicate a positive feeling of a safe place or home. Nonetheless, future studies could deeply focus on the relation between voice and (weather) sounds to experiment with fitting voice modulations that mirror the context. In general, sound bears an opportunity to reinforce calming situations, as raindrops against the window were positively associated.

### 9.8.3 Limitations

Our study investigated just one potential use case of sound overlays and VAs. However, a further holistic investigation is needed that requires testing several use cases to thoroughly understand how to use sound in voice interaction. Nevertheless, we could observe positive reactions to our design.

We mainly focused on developing a design approach and examining the general feasibility of a basic concept. At this point, we did not include advanced methods to examine the discriminative quality of the voice within our sonic overlays. Hence, we expect room for improvement in this area. In future work, additional quantitative studies, e.g., asking participants to transcribe the speech of the VA afterward, and using established Quality of Experience measurements as applied in telecommunications engineering [309], might significantly optimize the discriminative quality.

The same holds for our insights into semantic mapping and sonic associations. In the tradition of explorative qualitative research [158], our study uncovers relations and suggests hypotheses without statistical validation. For instance, our study suggests that the mapping and sonic associations are more coherent, when the illustrating situation (e.g. "it is raining") refers to natural sounds. Future studies should evaluate our insights and implications quantitatively to gain validated results that either confirm our hypotheses or show further areas of improvement [158].

Furthermore, the examples we tested were not representing real-time weather conditions at the location of our participants, nor were they presented in a realistic situation, e.g., during time pressure or participants knowing they need to leave the house in the next 10 min. To provide more robust results, tests need to be investigated that resemble both more realistic situations and feature the actual outdoor weather situation. Finally, our test was based on a rudimentary prototype that was not implemented and run on an actual smart speaker. We think that rerunning our study in a realistic and practice-based context might reveal further design principles and limits of usability but also opportunities for more sonic design.

## 9.9 Conclusion

We presented a study that aims to investigate what designers can learn from sound design if they like to enrich the experience with Voice Assistants. Focusing on one of the most favorite use cases, we present a user-centered approach to designing sonic overlays that complement the vocal messages of Voice Assistants and contribute to its user experience. Specifically, we were interested in how sonification of data might enhance voice interaction by using iconic, associated, and abstract sounds, in the example of weather forecasts. Based on a prestudy with 48 participants, we constructed a sound library for creating soundscapes for seven weather conditions: sunny, cloudy, foggy, thunder, rainy, freezing, snowing. We further evaluated the resulting soundscapes in an interview study with 15 participants to learn more about the effects of underlying spoken information with complementing soundscapes. Our study revealed both positive and negative feedback from our interviewees, based on which we were able to elicit respective design implications. Our design approach aims to open the design space for further sonic investigations and designs enriching voice interaction.

# Part III

# Research Outcome

# 10 Discussion and Implications

This Chapter will summarize the main findings presented in Part II and conduct a comparative analysis. While current research provides preliminary guidelines for structuring conversational interactions based on human communication, there is a lack of insights on designing and organizing conversations in ambiguous practices or contexts. As follows, this work will analyze the elements of practices as entities and performances [297], with each study thoroughly investigating into specific practices. These included cooking (cf. Chapter 8), interconnected practices of keeping track of fresh food, checking the freshness of food in critical moments of decision (cf. Chapter 7), and morning and evening routines of household members in the bathroom (cf. Chapter 6). By discussing the empirical work and design in light of the following research questions, I will derive design implications and conceptual contributions accordingly:

- **RQ1** How might Voice Assistants become Co-performing Agents next to humans?
- **RQ2** How can we design for a conversational co-performance of practices?
- **RQ3** How can multimodal agents contribute to an engaging co-performance?

Chapter 10.1 will focus on the evolution of voice assistants into conversational agents, emphasizing the voice-first paradigm and consideration of contextual user expectations. Additionally, Chapter 10.2 will utilize Social Practice Theory and Co-Performance to propose a conceptual design of interactive resources that facilitate the negotiation of human knowledge, meaning, and experience. Finally, Chapter 10.3 will explore how affective and multimodal design layers can enhance engaging interactions and enrich human experiences. Overall, all chapters equally address the three research questions, as the combined implications discussed throughout the chapters contribute to the holistic design of conversational agents. This chapter will then conclude by summarizing the contributions of this work and discussing the limitations and potential future directions in Chapter 11.

## 10.1 Envisioning Conversational Agents

Technological advancements and the perceptions and expectations of users shape the design space for conversational agents. Hence, this chapter delves into human home practices and aims to answer in particular *RQ1 How might Voice Assistants become Co-performing Agents next to humans?* Firstly, implications for voice-first interactions will be presented, followed by discussions on the contextual design that fits and supports human practices.

### 10.1.1 Towards Voice-first Interaction

**10.1.1.1 Contextualizing Information and Communication** The few established guidelines and experience in voice interaction design [60, 219, 42, 233, 220] contributes to designers adapting familiar interaction patterns, as the imitation of human conversations [60], or the use of visual information structures. Previous research suggests that using the term *conversation* to describe interactions between humans and smart home speakers may be inaccurate, as users tend to perceive their relationships with these devices as utilitarian or transactional and less casual [61, 254, 207, 332, 264].

The findings illustrate that intentional and proactive conversations initiated by agents typically aim to communicate requested information or provide support. Furthermore, current conversational interactions often either present a single piece of information or result in long monologues of data being read out. The designs of Chapter 6, 7, and 8 present how various contexts can provide opportunities for more or less casual exchanges of information, with the ultimate goal of prolonging the conversation until the user is satisfied with the knowledge retrieval. It is worth noting that casual conversations may not be the primary focus for users at present, as the studies (cf. Chapter 6, 7, 8, and 9) revealed a preference for utilitarian and purposeful conversational interactions. Despite functional goals, the studies show the benefits to incorporate communicative and conversational elements, such as conversational turns, information density, phrase formulation, active listening, and mutual understanding, without necessarily labeling them as true conversations in the traditional human sense. Based on the findings presented in Chapter 5, it is evident that most commercially available applications in CUI design do not prioritize conversations, but rather focus on providing access to information. This lack of conversational design should not be attributed entirely to unattainable design goals or user reluctance but to current limitations in usability and design approaches. Finally, conversations, in their true essence, involve meaningful exchanges of opinions, recommendations, activities, or ideas that are not solely driven by a specific purpose.

Moreover, this research aligns with previous studies on human communication [344, 16, 353] to distinguish between informative and communicative behavior. This can serve as a valuable guide for future design choices in effectively incorporating principles of human communication. Informative behavior aims to convey information without any subjective interpretation, although it can still elicit a response from the listener. On the other hand, communicative behavior specifically intends to evoke a reaction or serve a purpose for the recipient [353].

Study in Chapter 6 revealed that intentional communication of information involves anticipating and attributing meaning to the information as, for example, the recommendation to take an umbrella when it is going to rain. However, embedding recommendations in the conversation can sometimes evoke feelings of paternalism in users, leading to resistance to following the advice. On the other hand, as demonstrated in study in Chapter 9, providing a rich description of the information value, such as sonificating the weather forecast, allows users to gain an accurate understanding without feeling any external pressure to take specific actions. Current design practices for smart speakers prioritize providing general information to a broad audience to avoid the potential risks of offering unsolicited recommendations. For instance, the study presented in Chapter 5 shows that, when commercial voice assistants read out search entries, such as those from Wikipedia, they offer factual information but lack the ability to engage in genuine communication according to our participants. Their informative behavior is a unilateral process [344], that treats users as an audience, missing the opportunity to initiate a meaningful interaction. Shifting the behavior to communication would allow a series of message exchanges between the conversational agent and the human. At the same time, this interactive behavior would enable a sense of community that promotes the sharing of information and practices, with the chance of mutual understanding.

Further, effective communication necessitates the ability to adapt instantly to the interlocutor and the context of the conversation [16]. As presented in Chapter 7, research on the communicative behavior of experts has revealed their tendency to tailor information to match the knowledge level of their conversation partner. Experts employ specific vocabulary, gestures, and examples to clarify information and address questions when discussing different types of fish. They also adapt their conversational strategies and provide step-by-step guidance in response to specific queries from the recipient. This work's findings demonstrate the successful adaptation of these experts, while we were employing precise yet accessible vocabulary. Besides, investigations of the use of sonification of weather data indicate that adding means of expressing emotions and sensory impressions may support the subconscious interpretation by the users (cf. Chapter 9).

Consequently, to achieve meaningful conversations or interactions, design initiatives must incorporate communication strategies that prioritize contextualizing information and providing personalized recommendations. Without these considerations, as this thesis' findings demonstrate, the relationship between the conversational agent and human remains devoid of significance [53, 246].

**10.1.1.2 Voice-first Navigation** Envisioning user-centered conversational agents means taking voice-first and guiding interactions seriously while engaging in communicative behavior. Although users may initially struggle to adapt to these new interaction patterns, they are generally curious and open to engaging with emerging technologies. Transitioning towards prioritizing the auditory experience requires overcoming existing concepts that cater primarily to visual, screen-based interaction (cf. Chapter 5). By comparing participant statements, we discovered that their familiarity with screen-based applications hindered smooth interaction with the voice-based version. The participants assumed to use the same functions as provided by the traditional WIMP (Windows, Icons, Menus, Pointer) structures. Yet, the implemented design was not compliant with the auditory representation of data and navigation through voice and missed previously offered functions. For instance, user-generated recipe content often deviates from the standardized structures in contrast to professional providers, requiring explicit formatting and cleaning. Consequently, designers must reconsider current data models, interaction flows, functions, and services to ensure their adaptability and transferability to multimodal devices. This kind of approach would facilitate a seamless transition between different content representations. Additionally, providers may need to determine the range of services available based on the limitations of multimodal implementation.

The voice-first approach also emphasizes the need for ubiquitous access to essential information, as the study of Chapter 8 demonstrates: The CA, Cookie, is always on demand to support requests via voice, e.g., explaining the next cooking steps or switching on the oven. That design approach has proven effective in highlighting CA's role as a central point of contact and an orchestrator of various components, functions, and IoT devices. The evaluation results presented in Chapter 6 indicate the relevance of having a prominent assistant who can fulfill different roles, such as assisting children with brushing their teeth or providing up-to-date fitness tracking information. Users could access the mirror in the bathroom from various positions, with only a few applications requiring them to stand directly in front of the screen. Visual input was only an option when anticipating them to locate in front of the mirror because their performance required them to. In total, the evaluations confirmed that user expectations are met when the CA is responsible for user guidance and orchestrating interactions. Moreover, these findings apply equally to mobile application designers, although we did not specifically study the additional requirements of such devices.

Emphasizing a *voice-first* design approach, future interaction concepts and implementations need to ground applications in practice-based visions of users. While conversational agents serve as primary contacts to provide information or orchestrate functions, my findings suggest that power users of previously existing services, or those who have strong brand connections or product experiences, may take longer to adapt to voice-first services. In this regard, the complexity of an application should align with both the user's practice and their envisioned task goal, utilizing the combination of available modalities effectively [239]. Chapter 10.3 will further explore the implications of multimodal design relative to voice-first concepts, specifically examining the representation of visual information and the potential for proactive behavior by conversational agents.

**10.1.1.3 From Communication to Interaction** Providing a large set of skills requires efficient and effective management of interactive resources. One frustration experienced by users is the coordination of these skills, especially when switching between third-party provider skills or losing track of which skill is currently activated (cf. Chapter 5). For instance, there may be multiple shopping list applications available, but users can only use each of them independently, although not in combination with all skills. The study participants often found it challenging to stay oriented and differentiate between actions triggered by one skill or the CA. In comparison to supper apps [51], users faced challenges in effectively connecting their speech commands or outcomes to the actions performed by conversational agents. The nature of speech as a single auditory stream of information [88] contributed to a loss of orientation and awareness of the scope of available actions within the skill. Moreover, inconsistencies resulted in constant confusion for users and performance issues for the agent [220] using universal commands, such as "undo" or "skip" (cf. Chapter 5). Overall, users failed to understand the contextual hierarchy of global and local commands and demanded explanations for the cause of interaction issues.

Future conversational designs should put more effort into standardization for seamless navigation across platforms and services, including 3$^{rd}$-party providers that might contribute to fewer errors and unintentional closing of skills. Previous research [289, 68, 197, 196] suggested that users tend to underutilize the full range of functions offered by CAs, possibly due to a lack of personalization [53]. Appropriate countermeasures to further customize the CAs to user needs might be to choose their own activation words [53] or personal vocabulary and neologisms. If options for customization are given, the design should support the visibility and reconfiguration of skills. At least none of the voice interaction experienced participants in these studies were aware of such settings. Furthermore, the results can be generalized to similar eco-systems like Google Assist, as they might share the same causes of problems when integrating native and 3$^{rd}$-party applications. However, we can only speculate about language and word specifications in the case of, e.g., commands and activation words. The design case studies of this thesis (cf. Chapter 6 to 8), particularly aimed to address these issues by exploring conversational design and multimodal interaction to improve orchestration, resulting in overall positive evaluations of the design approach.

### 10.1.2 Fitting into context and practice

**10.1.2.1 Understanding at home practices** Prioritizing voice-first interactions and engaging in effective communication will result in user-centered CAs. Both humans and machines need to understand the cognitive and social rules of language usage [16, 59]. Therefore, the future design of CAs should align with social practices, norms, and the socio-material context of conversations and actions [297, 281]. Devices should not only be programmed and trained in NLP but also have access to a contextual understanding of these practices. In contrast to emerging sensations, humans learn to communicate and express thoughts and feelings throughout their lives, engaging in a complex process of participation rather than simply performing actions [344, 26]. As follows, the comparison of individual performances of practices in the work of Chapter 5, 6, 7, and 8 provides a comparison of the different elements of practices that represent ambiguous practices or contexts [297] to inform the design and content organization of CAs.

For instance, the choice of materials and artifacts in the design of conversational agents revealed what captures human attention within the performance of a practice. In the case of the bathroom, the mirror was chosen as the appropriate physical object to embed the assistant in the natural environment. The bathroom is seen as a private space for relaxation, hygiene, and personal well-being, which influences the function and role of the assistant. In contrast, the kitchen space typically requires more functional and productive support. The urgency and risks associated with different practices also vary. For example, checking the edibility of freshness can be reduced to the decision *to prepare to eat or not to eat* but requires solid competences with a higher risk of consequences of well-being than executing facial yoga to relax facial muscles and the mind, if not done properly at the first attempt. Further, the need for immediate information in meal preparation is higher than in post-exercise showering. Therefore, risks and urgency are evaluated differently in the performance of practice but could lead to positive or negative effects equally in the long-term for personal well-being.

The comparison of materials also highlights different approaches to utilizing CAs. While both home spaces allow for hands-free interaction, food is a material for physical consumption that requires attention but not ongoing physical interaction after the performance or consumption. This conclusion suggests that the materialization of the assistant differs from a bathroom mirror or a kitchen device. Nevertheless, both assistants contribute to building individual competences and supporting the personal and social meaning of a practice. Ultimately, the findings indicate that these considerations lead to meaningful and engaging interactions and experiences at home.

Unlike the externally controlled and structured work life of humans, households do not revolve exclusively around *getting things done*. As observed in the diary study (cf. Chapter 6), they spend their days engaging in various activities, both purposeful and leisurely, such as doing or thinking nothing, scrolling on the phone, or preparing for a hobby. These findings align with previous research [310, 148, 3, 310, 291, 5, 75, 128] that emphasizes the significance of smart home interventions and designs that cater to the diverse needs and processes of individuals. In opposite to early technology research that investigated and judged everyday life at home by workplace methodologies [69, 126], this work proposes to rethink the users' meaning of productivity and efficiency at home [76, 69].

Efficiency emerges from automating tedious manual tasks and monitoring unusual events initiated by the home or technology. However, the findings stress to discern enjoyable or meaningful activities that contribute to experiences of autonomy, competence, or pure joy, even if outsiders perceive these as less valuable or time-consuming, e.g., searching for restaurants, preparing meals, or grocery shopping. Some individuals prefer to stay in charge of control and decision-making while performing these practices, which adds to their sense of well-being. Therefore, it is necessary to limit autonomous notifications or distractions that can cause stress (cf. Chapter 6 and 8), such as work messages in the morning or ambient information about tasks and technology warnings. These attention-seeking interaction patterns are the opposite of actively supporting moments that often go unnoticed by technology and remain unconscious to individuals.

Further, this thesis highlights the significance of considering the learning capabilities and experiences of individuals when designing solutions for tasks such as determining the freshness of food. While a purely efficient and productive approach may provide a quick solution for users, it overlooks their own reconfiguration of practices by appropriating skills and competences. Presenting them with the final outcome would take away the deliberate sense-making. Previous studies [133, 211], including my own, have shown that users are concerned about becoming passive and unproductive in overly automated homes. Productivity is often associated with defined processes, milestones, and tangible outcomes, whereas being active implies engaged participation and conscious experiences. Experiences relate equally to entertaining bodily sensations and imagination aside from physical performances. For example, the morning use of sound-enhanced interactions with Alexa, for example, may take longer, but the participants have found that they experience a pleasant start and enhance their perception of the anticipated weather conditions by vividly engaging their imagination (cf. Chapter 9).

As a result, it may be beneficial to focus on designing artifacts and interfaces that promote and facilitate shared engagement and performance in activities. These designs should also enable a meaningful redistribution of time and space, empowering users in their actions. For instance, different areas within the home can be designated for different types of tasks, ranging from coordination and communication work to creating a calm and mindful environment for interactions.

**10.1.2.2 Contextualizing Skills of Conversational Agents** Through the in-depth study of home practices and context, we have identified opportunities to enhance the capabilities of CAs and aligned them with user practices. Similar to previous research, this work's findings indicate that while users are interested in a range of skills and interactive support offered by CAs, they are often disappointed by the implementation and interaction design of these features (cf. Chapter 5). Compared to the services from websites or smartphones, participants expected the same quality and scope of services, e.g., access to specific recipes or functions like loyalty programs. Thereby, the missing depth of skill implementation annoys users because it suggests the principal availability and facilitation of skills as well as its match to the household practices but is not sufficiently implemented to benefit users. When users attempt to accomplish tasks and goals, they encounter usability and interaction issues that further hinder the usefulness and overall experience with CAs.

Unfortunately, despite promises of commercials and the impressions evoked by the availability of endless 3rd-party provider offers, current CA platforms offer only some reliable skills with limited options to mix and match those. Hence, the concept of supper apps [51] highlights the significant advantage of an integrated solution. The findings of this thesis demonstrated the benefits for users when skills align with specific use cases, user routines, and contexts. The communication of the scope of skills should reflect the actual skill set. In the future, however, these should become extended to ensure seamless interoperability and customization by the user. Although skills do not currently generate natural conversations, they satisfy utilitarian requirements at some level [61]. With further advances in skill design, the ecosystem holds the potential to evolve from a knowledgeable but shallow CA [120] to becoming smart, useful and engaging.

**10.1.2.3 Becoming a Conversational Agent** Following the design principles by supper apps [51], users perceive CAs as seamlessly managing all requests. This perception is created by the design of a single voice that handles all communication and actions. Therefore, aligning with conversations [61] and human practices means meeting the user expectations regarding the role and responsibilities of the agent. As discussed in the first subchapters, the use of speech and voice-first design evokes the impression of intelligent behavior

[333, 332, 119, 53, 289], which directly influences the envisioned scope of behavior, interaction, outcomes, and capabilities or skills, as shown by this work's data. The assignment of autonomous behavior leads participants to describe the conversational agent as somewhere between a tool and an agent, as implied by the term itself (cf. Chapter7). Therefore, the effective design of conversational agents involves more than just controlling actions; it also requires considerations of the assistant's role and personification to enhance user interaction and experience. In the study of Chapter 8, participants appreciated the design of Cookie and its human-like qualities. They mentioned supportive conversations and feeling less alone while cooking. Additionally, conversational personifications create a motivating atmosphere by providing personal praise and validation for users' actions (cf. Chapter 6 and 8). Participants also assigned teaching and coaching functions to the assistant and relied on its credibility as a collector and recommender of information, e.g., trust in the origins of information (cf. Chapter 7). However, participants discussed the need for different teaching styles that align with users' learning preferences. Chapter 8 derived four personality types that may influence the acceptance and interaction between humans and conversational agents, which will be further discussed in Chapter 10.2.2.4.

Users evaluate the benefit of conversational systems based on utilitarian factors, while they likewise value the social presence of the assistant. The assistant doesn't need to imitate a specific character but meet user expectations in terms of support, such as tone of voice, explanations, and proactive recommendations. In line with these expectations, the deliberate design of informative and communicative behavior [344, 353, 15] benefits both the coordination of skills and the direct interaction with the user. In summary, the design of conversational interactions should be multi-component and multimodal, and it should align with the practices and personal challenges that users face during their interactions.

## 10.2 Engaging with Conversational Agents in Co-performance

Transitioning from communication to interaction with conversational agents, we delve into the second research question, which explores the design of guidance in action: *RQ 2 How can we design for a conversational co-performance of practices?* In this regard, the subsequent subchapter investigates the classification of conversational agents as carriers of practices [297] in co-performance [177] with humans. Both aspects contribute to the advancing of previously unknown or emerging practices and the required know-how in applying skills. The Sense-Think-Act cycle [250] is utilized to shape the interaction and assess the abilities and contributions of both parties. Ultimately, the question arises as to which level designers should consider human agency, autonomy, and competence.

### 10.2.1 Carriers of Practice

**10.2.1.1 Negotiation of Knowledge** Moments of crisis frequently demand new approaches to address issues that challenge the competences and skills of the concerned person. In such situations, individuals may turn to either material, such as tools, infrastructures, or manifestations of knowledge, or other people, such as knowledgeable relatives, friends, professionals, or lay teachers and coaches, who can provide trustworthy advice and guidance. The latter approach not only provides access to information but also teaches how to apply that knowledge in practice. These carriers of practices shape and define the practices by constantly performing variations and blending them into society [297, 30, 183]. This thesis encloses and investigates examples of such practices, for example, cooking, assessing edibility, or maintaining dental hygiene. Over time, novices can become new carriers of practices or replace existing ones, introducing new variations and elements [297, 30, 183].

Hence, carriers of practice interactively support the acquisition of knowledge and the reconfiguration of practice elements. In this light, the question arises whether users perceive conversational agents as mere materials or carriers of practices. So far, HCI research in the field of Social Practice Theory views technology predominantly as tools or infrastructure facilitating the performance of practices and representing a manifestation of knowledge and socially shared understandings [333, 274, 320]. Materials, either physical or digital, are seen as objects that are utilized and controlled by humans.

Kuijer et al. suggest that "artificial performers should be considered as a category in their own right and not as (poor) imitations of humans ones." [177]. Conceptualizing CAs as carriers of practice acknowledges their ability to provide interactive resources for experiential learning and support novices to perform practices while appropriating knowledge. Unlike linear forms of information, such as text or speech, CAs possess artificial agency and can apply knowledge by adapting to specific practice situations. For instance, the designs of interactive cooking (cf. Chapter 8) or food assistants (cf. Chapter 7) offer diverging paths that align with the sayings and doings respectively decisions of the humans [281, 297]. Depending on the level of institutionalized knowledge [108], the CA actively listens and considers multiple alternative paths, evaluating situational information from the user to adapt static and linear information. Informative behavior [353, 17] aims to provide personalized knowledge to the recipient. By allowing for follow-up questions and actively participating in task execution, conversational agents serve as interactive counterparts in the performance of practical tasks (cf. Chapter 7 and 8).

Consequently, the communicative behavior [353, 17] of CAs represents their intelligence. They can adapt their expressions and explanations to match the proficiency level of the user, providing recommendations for cooking steps and tools or describing freshness levels. The conversational capabilities of the prototypes shown in Chapter 6, 7, and 8 are limited compared to current LLMs that offer even more extensive communicative adaptation. However, the capabilities of LLMs do not contradict the results above but rather open up new possibilities for user interaction and experience for voice-first approaches. The support provided by CAs is not dependent on human imitation and instead utilizes human-like qualities to guide users in their actions, similar to partnering with a coach, teacher, or like-minded individual (cf. Chapter 6, 7, and 8).

Furthermore, as discussed previously in Chapter 10.1, the social presence and speech-ability of CAs creates a sense of immediacy, directness, and clarity in communication (cf. Chapter 5, 6, 7, and 8). As research has shown, communication affects behavior [344] and leads to moment-by-moment adjustments in responses [16, 15]. Consequently, this capability holds significant potential for influencing and shaping human practices. The basic idea is that effective learning necessitates the presence of an individual with expertise in the specific practice to impart its complexities [297].

The circulation of knowledge goes beyond sharing information and builds upon the early negotiation of meaning, material, and competence [297]. The discussion of the elements leads to a (re)newed mutual understanding of the practices, encompassing not only the knowledge of what to do but also the social aspects of knowing-how, which are transformed and adapted in the new context [297].

Overall, the design of informative and communicative behavior exerted by conversational agents serves as interactive support to interpret and apply information. Further, understanding and applying knowledge presupposes successful negotiation with an interlocutor. As a result, a carrier of practice, such as a conversational agent, not only provides information but acts as an integral part of negotiating knowledge. Thus, to advance the design of CAs, we need to be aware of a social codification and reconfiguration of practices. Adapting and integrating social dynamics to the conversational interaction might overcome the simple mimicking of human-likeness [9, 68, 333].

**10.2.1.2 Negotiation of Meaning** Lave and Wenger [183] challenge the notion of individuals as foremost cognitive beings, emphasizing the significance of personal and social factors in knowledge, skills, tasks, activities, and learning. Accordingly, humans need to negotiate meaning to learn to make sense of the world and to engage with their surroundings effectively [350]. Social negotiation of meaning occurs notably through the sharing of knowledge and experience between current practitioners and novices who introduce new perspectives and past experiences to the community [350]. Similarly, this research (cf. Chapter 6, 7, and 8) demonstrates that individuals engage in the negotiation of meaning when interacting with conversational agents, particularly when encountering unfamiliar or challenging practices. For instance, during the exploration of the smart mirror application for children's teeth brushing, participants expressed differing opinions. Parents discussed whether it is appropriate to relinquish control over oral hygiene to a CA to support the children's autonomy. This interaction with the CA prompted a negotiation of meaning regarding an established household practice and parents' responsibility and care. As revealed by Chapter 7, the cooking assistant potentially impacts the meaning of meal preparation. Currently, the design primarily focuses on learning new recipes. Future approaches might, for example, ensure user goals to archive meaningful family recipes for regular preparation following the family's traditions.

In conclusion, CAs can challenge or reinforce the meaning of practice, but the negotiation of meaning remains primarily one-sided with the human participants. Currently, CAs serve as mediators that enhance the meaningful engagement of humans but do not independently generate meaning for themselves. In contrast to social settings, Becker's research [18] on marijuana users shows how novices undergo a transformation into experts and redefine their identities through meaningful interactions and practices within their community [297, 18]. While CAs contribute to the human process of learning, evolving, and belonging to a CoP, they are not members themselves. They take on roles and behaviors assigned by users or designers but do not craft their own identities in the moments of negotiation.

Studies indicate that the relationships formed within communities have a profound impact on the understanding, retention, and performance of practices, as well as on individual identity and meaning [183]. CoPs are a renowned example of this, representing a natural and effective approach to learning [349, 183]. These communities consist of individuals who share a common interest or passion for a specific activity and continuously enhance their skills through regular interaction. Often, this learning process occurs organically, without a deliberate intention to establish and share a collective practice.

While CAs promote the sense-making and meaning-making of humans, in the studies of this thesis, the participants did not treat or refer to them as an active part of a CoP. Unlike human participants in a CoP, CAs do not actively seek acknowledgment or strive for a sense of belonging within a community. Consequently, they do not create new knowledge through interactions with members to legitimize their membership. However, future learning algorithms that incorporate user feedback and adapt their knowledge accordingly could enable CAs to evolve in this direction. Still, it seems unlikely that the collective sense-making that characterizes a CoP will be fully realized in CAs. First, they have to demonstrate a genuine interest in developing socially significant practices to build identity and select values to participate in and preserve the community.

With CoP following the idea of situated learning, the processes of reflection and active engagement within a practice can serve as valuable guidelines for design. The mere proposition of interactive knowledge, on the other hand, is not sufficient to fully replicate the experience of participating in a community. As such, individuals require the intrinsic motivation to solve problems collectively or to independently seek original experiences and knowledge that align with the values and meaning of the community, for example, individuals who share a commitment to reducing food waste (cf. Chapter 7) or promote a healthy lifestyle (cf. Chapter 6). CAs can assist in negotiating meaning, even if they are not considered an active part of a CoP, by embracing roles such as inspiration, classification, or institutional leadership.

Recognizing CAs in the roles of carriers of practices [297] that contribute to the reconfiguration of practices turns the design space to concepts that align with the current perceptions of users. As discussed earlier, users tend to perceive conversational agents as intelligent counterparts capable of providing support beyond a simple tool, such as a pan or smart oven, by engaging in human-like conversations. Notably, users stay in the position to decide what is meaningful to them while changing their opinions seamlessly whether CAs are a tool or agent [274]. Nonetheless, CAs as carriers enable meaningful and engaging interactions. The nature of conversational interactions requires adaptable and personalized infrastructures and tools that provide a sense of social presence, motivating and comforting users. With the design process still in its early stages, Conversational Designers could use these insights to study and prototype materials to enhance knowledge transfer and transform that into actionable advice. In conclusion, by applying the Social Practice Theory [297, 262], we can effectively evaluate the abilities and significance of conversational agents and identify potential design opportunities.

### 10.2.2 Designing for the Co-performance of Practices

Following the idea that CAs can function as carriers beyond basic materials, this chapter will discuss the design space of CAs as actors in co-performance with humans. The concept of co-performance [177] enables us to assess the distribution, balance, and interplay of human and non-human capabilities and their contribution to learning by experience [177, 110, 111, 163]. Furthermore, speech represents a decisive design feature to align sayings and doings in practice [281].

**10.2.2.1 Towards the Negotiation of Experience** Lave and Wenger [183] argue that apprenticeship does not necessarily follow a strict hierarchical structure between a master and an apprentice. Instead, they suggest that anyone can share knowledge and contribute to the learning process, creating engaging opportunities for learning across different contexts and timeframes. This process involves a dialectical negotiation of knowledge, experience, and decision-making for past and new knowledge to circulate and evolve into practice. Previous studies demonstrated great potentials to use CAs for educational and companionship purposes [102, 200, 71, 137] but without proposing specific design implications to create negotiation processes with CAs [137].

In this light, this thesis serves as a starting point to address the design space of negotiation processes between CAs and users. For example, require users, when they encounter critical moments of food consumption, the negotiation of embodied and institutionalized knowledge, e.g., assessing the freshness of food or choosing preparation methods for meals [134]. The findings suggest that by its interactive nature, the agent is capable of guiding humans to experience the necessary considerations to make informed decisions (cf. Chapter 7 and 8). By not following advice or instructions, the participants still experienced moments of reflection, where they had the chance to evaluate their behavior later without blaming the agent for not having prevented false decisions. Moreover, food professionals emphasized the significance of embodied experience to make effective decisions in the moment and the future. As for brushing teeth or face yoga (cf. Chapter 6), participants appreciated formalized knowledge but emphasized the advantage of following body movements. The general supportive and interactive design of the prototypes did not force or center around learning but aimed to solve situated problems right away, similar to the observations in CoP. The optional support of CAs never felt patronizing or overruling human decision-making (cf. Chapter 7). Despite that, participants expressed skepticism about the trustworthiness of the advice given by the CAs and suggested the addition of sensors to validate their personally experienced and embodied sensations.

Finally, the design should promote humans' trust in their senses, the classification of sensations, and the development of embodied knowledge. Future CAs should aim to motivate, guide, and inspire individuals to perform practices independently from technological support.

**10.2.2.2 Carriers of Practices in Co-Performance** In line with the results of Kuijer et al. [177], this work shows that sensing capabilities of artifacts and agents enforce their autonomy and the impact on human decision-making. This highlights further the disparity in embodied knowledge and experience between humans and smart artifacts. The notion of co-performance emphasizes the active negotiation process between humans and technology to determine the experience and meaning of a practice.

Using the Sense-Think-Act cycle [250] as a guiding framework facilitates the systematic distribution of tasks, steps, and capabilities in co-performance on three levels. This framework emphasizes that intelligent machines must first utilize their sensors to gather information, then process it through computation, and finally take appropriate actions within their immediate surroundings and context [250]. While researchers commonly employ it to analyze machine intelligence from a human perspective, this thesis stresses that actual intelligence and agency emerge from the collaborative interaction between humans and technology. Further, this framework sensitizes designers to the potential challenges of mismatching capabilities.

Specifically, individuals with prior cooking or food experience emphasized the significance of personal experience and the empowerment of their senses. This observation exemplifies the tension field of efficient and accurate decision-making that either experts or lay people can leverage by just using their embodied tools to get to a satisfying conclusion. Hence, an effective and empowering design does not always involve additional or cutting-edge sensors, such as verifying food edibility, but instead builds on co-performance with CAs to yield results through deliberate and autonomous human action. Simultaneously, this approach takes conscious interaction seriously [133] to enhance the overall experience and human engagement in performing practices.

The studies in this thesis revealed that the CA serves as a tool for learning, temporarily assisting users in their decision-making process by engaging their own thinking and sensing to classify their perceptions and performances. From this perspective, CAs intervene to actively support reflection by demonstrating practices, which aids in the transition from relying on institutionalized knowledge to developing embodied knowledge. The results in Chapter 7 show that the prototype, Fischer Fritz, effectively addresses this issue by offering a systematic, step-by-step approach that instills confidence in users' actions. Finally, the work of Gherardi and Nicolini [108, 109] emphasizes that 'competence-to-act' is grounded in knowledge. During co-performance, the conversational agent serves as a mediator, revealing the connection between specific actions and know-how for the user.

On the other hand, participants in a smart kitchen still appreciated the autonomous support of smart appliances such as the stove or oven (cf. Chapter 8). The findings point to the experience of competence and multitasking, as setting timers or pre-heating the oven is not an overly demanding or meaningful task, and users are eager to delegate those to technology. The time pressure of preparing meals requires to manage resources strategically. In terms of cognitive processes, users often struggle to make sense of their bodily responses to different materials, e.g., food, leading them to rely on institutionalized knowledge [108], such as shelf life and food disposal practices.

Lastly, the findings align with the principles of purposive learning and active engagement in practice [109, 354]. While the CA may assist in certain aspects of thinking, intellectual growth and training unfold through self-reflection and the exchange of embodied knowledge, which relies on effective human decision-making that leads to a sense of self-competence and autonomy. Providing continuous guidance and mentoring in performing actions helps individuals develop the ability to act independently and establish new practices over time.

**10.2.2.3 The Role of Speech Co-Performance** Within co-performance [177], speech anchors the sensations and actions of the human and the CA when communication between them is functioning. The evaluation study in Chapter 7 revealed that users did not assess Fischer Fritz based on its resemblance to a human but on its technological capabilities. Therefore, the following studies (cf. Chapter 8 and 9 ) investigated language as a material to incorporate human characteristics, such as empathy expressed through specific words and sounds, but still tie it to technological capabilities and not to mimic human understanding and relationships. As discussed in Chapter 10.1.2.3, the results of Chapter 8 confirm the words and expressions of the CA created a motivating atmosphere.

In regular human CA interactions, users give direct commands to the machine and act through it [289]. The case studies in this thesis highlight a different approach when both the user and the CA actively engage in the real world through speaking and listening, co-performing in tandem. Users saw the agent as responsible for fulfilling its purpose by providing informative and understandable explanations and guidance tailored to the user's abilities. The findings of the preliminary study in Chapter 7 suggest that the language used in the specific task of assessing fish relies heavily on metaphors and figurative language. The evaluation demonstrated the challenge of balancing concise commands and clear instructions that move the co-performance forward without overwhelming or confusing users and how mutual reliance and a common language enable accomplishments beyond executing simple tasks [289]. Therefore, to enhance the coaching aspect of the CA, it is necessary to ask open-ended questions, accommodate for mistakes, and encourage exploration. This approach to dialogue entails incorporating intelligent fallback options that do not result in dead-end conversations but instead provide enlightening and encouraging responses [197, 53], that we were able to implement in the cooking assistant of Chapter 8.

This separation of human sense-making and machine thinking requires a common language. Previous attempts at utilizing voice assistants have not effectively involved humans in purposeful collaboration or meaningful conversations [268, 53]. Co-performance serves as a model to link the sayings and doings of the carriers to result in mutual practice and for the human to evolve competence. Thereby, conversational abilities and proactive behavior contribute to a perceived intelligence that does not feel endangering but supportive and engaging. While humans naturally rely on their senses, they require the agent's conversational guidance to interpret sensory information. Thus, their distributed capabilities complement each other.

**10.2.2.4 Balancing Agency and Proactivity in Co-Performance** Prior work highlighted the benefits of shifting research from proactive computing to proactive people, emphasizing the design of engaging artifacts that facilitate meaningful interactions and collaborative decision-making [268]. Past chapters have argued how voice assistants might become conversational agents by co-performing next to humans. User-centered guidance and valuable demonstration of practices require CAs to anticipate likely situations [232, 173] and take the lead from time to time to proactively make suggestions for alternative pathways. Despite or because of the lack of CAs' accurate modeling of users' performances, emotions, and intentions [268, 232], we need to balance both the human and non-human agency in their co-performance to avoid technological over-dependency and paternalism. Proactive behavior of CAs can either prevent mistakes and lead to success [173], yet some people feel that technology is questioning their autonomy and competence (cf. Chapter 8). Instead, designing and implementing adaptive levels of proactivity contribute to the flexible and successful matching of users' expectations and perceptions [365], leading to an equal distribution of tasks along the dimensions of competence, senses, and agency between both human and non-human actors (cf. Chapter 7 and 8). In line with existing research [232, 172, 365], Chapter 8 proposed levels of proactivity to match users' personalities and current situations. While further research is required to create a reliable instrument for developing technology based on generalized cooking types, this study serves as an initial step in raising awareness of the significant variations in proactive design.

In more detail, the design case study in Chapter 8 illustrated an approach to value human autonomy and competence as design parameters to match proactive people with proactive behaviors of CAs transparently. The findings revealed that the user groups demonstrated varying needs for trust, autonomy, and competence that influenced the likelihood of accepting the recommendations by the assistant. Those with a strong desire for autonomy profited from error prevention measures and offering several exit points along the process, while those with high levels of competence welcomed challenges that encouraged more creative approaches. Beginner users tended to follow Cookie's suggestions and instructions closely, viewing the provided information as an opportunity to gain knowledge and know-how. In contrast, accurate cooks were free to explore alternative paths and choices, with the potential for support in more advanced recipes or unfamiliar areas. Creative cooks who valued autonomy expected proactivity to prevent mistakes. In general, all users considered control to activate or deactivate proactive features of the assistance.

The design challenge of user-centered proactivity arises from the tension of the humans to embrace proactive CAs that support their practices and decision-making but simultaneously hesitate to leave all control to them [57]. In Chapter 8, users felt a loss of control when appliances operated autonomously without their consent, particularly in situations involving safety or complexity, such as self-operating of the stove. Providing users with explicit choices and settings for each appliance and function, such as manual control, temperature regulation, or last-minute user interventions, can mitigate the loss of control and enable a reasonable sharing of competences and actions from the perspective of the user (cf. Chapters 7 and 8). On the other hand, automation may ensure consistent and efficient meal preparation in fast food processing [217]. Users particularly benefit from these functions in high-pressure or time-sensitive cooking situations when striving for reliable and repetitive results. In line with self-reliance and situated learning [183], this cooking assistance has the potential to enhance individuals' understanding of heat control as a fundamental aspect of meal preparation and potentially enhance current human food practices for the future [3].

Meanwhile, the smart mirror did not exhibit any proactive behavior besides the synchronizing information but rather an informative and communicative behavior (cf. Chapter 6). Nonetheless, proactive reminders like taking an umbrella in case of bad weather caused users feelings of patronizing by technology. In contrast, the design of Cookie (cf. Chapter 8) revealed that reminders and suggestions elicit higher levels of trust in users than uncommented or not communicated autonomous intervention [173]. In general, users appreciated these moderate levels of proactivity.

When using proactivity as design material, we have to anticipate limitations in fully automated systems, such as speech recognition and user activity detection, as they may result in lower levels of trust. In this light, design has to account for users' different needs and levels of trust, autonomy, and competence. Flexible and, particularly, moderate levels of proactivity increase human's perceived control and support their self-reliance.

Creating conversational agents that "act" [329] as interactive resources can establish a social presence, fostering relationship building [151] and long-term trust with users [53], as the findings in this work confirm (cf. Chapters 6, 7, and 8). Chapter 8 demonstrated users' perceptions of their autonomy and competence significantly influenced their overall satisfaction and enjoyment of the experience with the conversational agent [131]. All presented CAs in this thesis revealed that maintaining a balance of agency plays a significant role in fostering a positive attitude towards technology and facilitating the acquisition of new skills (cf. Chapter 6, 7, and 8).

Dolejšová & Wilde et al. [79], have cautioned against an excessive focus on automation and technology-centered designs, as it may disconnect users from the social and cultural significance of practices, e.g., food. Prior research [310, 291, 5, 75] emphasizes prioritizing user needs and avoiding excessive technology dependence in alternative visions of smart homes. Examining the evaluations of the CAs in this thesis (cf. Chapter 6, 7, and 8), participants highlighted the sensory experience-making in the performance of practices. Thereby, they appreciated a conscious experience of materials, such as food, the mirror, their surroundings, or their own body. Although some participants of Chapter 7 remarked that sensors would provide additional trust in decision-making, most of them highly valued the integration and emphasis on the human senses. Some mentioned that extra sensors would impede human sensory training and, thus, agency. Further, the findings in Chapter 8 indicate that future designs should include adaptability and flexibility to provide an over-write function to save personal variations that can be, for example, shared with others. This approach emphasizes the importance of recipes, creativity, and self-expression in cooking, rather than striving for a *perfect* but standardized version through automatized instructions by a CA [79].

Here as well, different levels of proactivity might regulate the experience-making and should acknowledge the competence and autonomy levels of its users. While CAs prioritize user competence and autonomy, they should also address the unique challenges in practices, e.g., procedures, choice of materials, and timely information. Users should be able to modify and instruct the agent to either facilitate faster decision-making or provide additional information to enhance knowledge and learning. By pushing the agency step-wise towards the human, promoting learning, trying, and self-performance, technology reliance could be minimized to technology none-use and the performance transformed into human practice. In contrast to full automation, the active participation of humans in the decision-making process promotes a balance of control and proactivity.

In summary, we should pay particular attention to determining the appropriate level of proactivity and the accountability associated with the decision. Adaptive solutions should aim to adapt to users' personalities and contexts without overly constraining or controlling the situation. These design decisions will impact the risk and outcome of failure, the meaningfulness of the performance and practice to the human, and their levels of competence and need for autonomy. By balancing the amount of IoT interventions, users can regain their autonomy and feel empowered when, for example, cooking for themselves and others.

## 10.3 Enriching Conversational Interactions and Experiences

With voice-first designs being still at the beginning of the interaction design evolution, multimodal approaches might mitigate some of the limitations of voice-only [197, 221, 351] and contribute to an engaging experience that integrates narrative design parameters to evolve speech-based interaction [292, 9]. When interactions unfold, they evoke certain perceptions, emotions, and experiences in users that impact the quality of relationship and communication between both parties [95]. This thesis focuses on the experiences that emerge in interaction and co-performance with CAs while integrating the embodied knowledge of humans, immediate experience with its impact on competence, meaning, and material, and, thus, the situated and socially embedded context of practices. CAs as interactive systems aim to provide engaging experiences in the co-performance of practices, which means to decide from a design perspective "how best to represent and present information that is accessible via different surfaces, devices and tools for the activity at hand." [268]. Further, the utilization of multimodal interfaces enables the expression and implementation of CAs' proactive behavior, in addition to considering various levels of proactivity (cf. Chapter 10.2). Multimodality is typically understood as the sum of individual modalities. Within the scope of this thesis, the modalities are intended to enhance and supplement interactions with the CA, such as sound, graphic, and tangible interfaces. Discussing multimodal options of informative behavior and subsequent user interaction raises questions about the balance of communication and expression mechanisms to enrich interaction and experience: *RQ3: How can multimodal agents contribute to an engaging co-performance?*

### 10.3.1 Sonic Interfaces

The prevailing design approach uses sound to convey information and replace functions and representations [27], usually implemented in the form of auditory icons and earcons that are easily recognizable. However, this approach requires either a clear sonic representation or users to learn the meaning of the sounds. For example, earcons intend to signal warnings or draw attention to events [27, 107]. This focus on iconic sonification may limit the ability to create rich soundscapes that can evoke emotions, create atmospheres, and provide immersive experiences, as seen in traditional media and extended realities [152, 271, 46, 150]. Data sonification has a dual function to communicate information and express emotions [271]. Sound Design proved to be an effective tool in Science Fiction [352] as well as in games and XR [150] to create imaginary worlds or provide less common experiences.

The findings in Chapter 9 suggest that while sonification offers new possibilities for design expression [298, 49, 271], voice communication remains a precise and efficient way to convey information. For instance, acoustic feedback in the form of a beeping sound may be less meaningful to users compared to clear voice output that provides specific suggestions or warnings. Users expect sounds to align with the information provided through voice channels, avoiding contradictions. Additionally, related concepts or information, such as frost and snow conditions, need extra investigation into representation to prevent misunderstandings. A further challenge to represent through sound is probabilistic or numerical information like a 50% chance of rain. In this case, a scale encompasses numerical information that needs recognizable and recurring values that increase and decrease via sound. Therefore, combinations of speech and sound (cf. Chapter 9) enhance the informational and affective value of the weather forecast, leading to a further nuanced understanding of the message. While speech communicates unambiguous information, e.g., temperature or humidity descriptions, users understand the message effortlessly. Rich descriptors of colors, textures, or smells similarly contribute to an immediate and strong sensory sense-making (cf. Chapter 7), anchored in imagination and past experiences. However, the created sound overlays succeeded only when they fitted the cultural context and experiences of personal residence environments. These have vastly impacted the subjective interpretation and association of the provided information, for instance, festivities like Christmas, or living environments close to nature versus large cities (cf. Chapter 9).

Moreover, the results show mixed opinions about the structure of the audio clips, specifically regarding the timing of the sound and voice (cf. Chapter 9), with some users preferring the sound first, followed by the voice, and ending with more sounds. They felt that this sequence enabled them to form an initial impression of the weather based on the sound, which was later confirmed and clarified by the voice. Other participants mentioned the clips were too long compared to a concise voice-only weather forecast. While most participants agreed that the sounds enhanced their connection to the weather compared to the voice alone, some suggest listening to the voice first for a quick retention of most information.

Expressing or inducing emotions might also contribute to meaningful experiences that enhance relationship building between the interactive counterparts [61, 33, 35, 178]. Therefore, the design approaches in this thesis further investigated creating the surrounding and ambient sound that enhances the emotional experience of CAs. In the realm of expressive and informative interaction, designers have a responsibility to consider the sonification of positive and negative experiences carefully. The results revealed concerns about the potential manipulation of sounds, particularly when discussing news (cf. Chapter 9). Some individuals will prefer to receive factual information without emotional embellishment. Additionally, some users intentionally avoid triggering negative emotions. Therefore, designers should strive for a balance by incorporating sounds that convey a sense of safety or comfort when, for example, dealing with hazardous weather conditions like thunder.

Designers should be cautious when using abstract concepts that involve human voices, as this may lead to confusion. However, a cohesive combination of illustrative sounds and real-time voice strengthens the perception of the message. Currently, the voice is simply talking over the soundscape after a few seconds. Using speech should aim at different levels to encode information. Besides purely functional messages, not only voice modulation [198, 347, 1, 169, 288, 286, 187, 37, 87] and sound but also the choice of vocabulary and idioms contributes to an informative and rich communication that facilitates immersive and engaging experiences. Future studies could delve deeper into the relationship between voice and sounds, experimenting with appropriate voice modulations that reflect the context. Overall, sound is an opportunity to enhance calming and positive situations, for example, listening to raindrops against a window. In general, further integration and interpretation of data, e.g., location, living environments, chronic data, and lived experiences in the area of users, allow designers to create auditory-based interactions that enable personalized conversations. Beyond choosing a representative modality, we need to prioritize and classify information linearly, similar to the prioritization levels for hints discussed in Chapter 8.5.2. Finally, users have to be informed about these levels.

By expanding the role of sound in voice interaction, the findings in Chapter 9 suggest that incorporating sonic overlays can enhance the encoding, illustration, and communication of messages. Using iconic, abstract, and symbolic sounds improved users' perception of weather reports through speech-based interaction and led to an engaging and enjoyable experience. Hence, iconic elements in the sounds helped make the intended messages more recognizable, while the absence of certain iconic sounds might obscure information, e.g., the representation of fog. Moreover, music elements and abstract soundscapes contribute to creating a comprehensive impression of specific weather conditions and conveyed moods and emotions.

In summary, the combination of sound and speech effectively enhances the intended message. Speech provides precise information, particularly for events or impressions that are naturally silent and difficult to represent through sound. Additionally, careful consideration of granularity and discrimination in sound design can improve the accuracy of information. However, the timing of sound and speech integration in the overall design requires purposeful consideration and further research to provide clear guidelines. The challenge presents itself in effectively combining abstract soundscapes that enhance the experience and iconic sounds that ensure the communication of messages.

### 10.3.2 Adding Graphical User Interfaces

Currently, IPAs like Alexa Show provide conversational and visual access to contextual information but lack effective interaction to guide users through information hierarchies (cf. Chapter 5). Unlike GUIs, users have limited control over speech-based systems due to their "invisible nature" [66], which complicates reviewing and modifying past actions or commands [295]. Also, commercial CAs with an attached screen, like Alexa Show, do not improve the discoverability of content and skills. Additionally, users faced difficulties when verbally editing long lists of items, as it took time to listen and process the information, leading to increased cognitive load [239] and potential frustration (cf. Chapter 5). Finally, users predominantly used touch to end frustrating conversations, as they could not operate skills by voice only or touch (cf. Chapter 5). Chapter 8 indicated how to successfully determine appropriate interaction modalities for assisted cooking in smart kitchens. After considering various options, including gestures, voice, and graphical interfaces, we decided that a GUI with touch and voice showed the most potential for effective interaction and instruction, based on previous discussions with users. In line with [345], the research outcomes in this thesis suggest that multimodal interaction can address the limitations of individual modalities, such as voice recognition in noisy environments or touch input with dirty hands. By offering multiple ways to interact, users can engage with the system using the most situated modality.

These observations reveal that proper mode selection can only occur when the specific needs and characteristics of a task are known and met. The differentiation between input and output modalities enables us to prioritize and express information accordingly. For example, these considerations apply particularly to mobile phones, as users expect them to serve multiple purposes at any time. However, further research into specific mobile applications may reveal use cases that necessitate explicit voice input, such as cooking or driving.

In line with the need for robust heuristics for CUIs [60, 220], this work suggests aligning skills with users' performances of practices and tasks, particularly when a switch in modality is involved, highlighting the significance of smooth transitions for users. The findings in this thesis (cf. Chapter 5) uncovered that once the participants used touch input on the Echo Show's display, they could neither resume prior conversations with a different skill and had to close it nor return to previously opened skills as they had to start the task all over again. Therefore, users need notifications when a CA skill requires switching the modality for proper functioning and advanced maintenance of the progress of opened applications. In this light, users would appreciate clear signifiers in different modalities to enable intuitive interaction.

Overall, the burden of interpretation and establishing mutual understanding was placed on the users, as the CA lacked proper feedback and effective integration of visual and auditory signifiers. The absence of both auditory and visual signifiers contributed to users' feeling a loss of control and limited the discoverability of skills, negatively impacting the participants' orientation and brand recognition. Creating dialogues and providing visual support to enhance comprehensibility and convey emotions presents an opportunity for design to compensate for the CAs' limitations in interpretation performance.

### 10.3.3 Interfaces embedded in Artifacts

Traditional designs of smart homes envision technology as concealed infrastructure that supports monitoring and controlling appliances, such as lights and music, news, or set timers and reminders [6, 268]. Likewise, smart displays and speakers operate as ubiquitous information hubs and access for control [60, 8, 7]. Further, designers and researchers conceal technology by embedding it into artifacts, like mirrors [60, 8, 7]. Instead of hiding purposes, we could leverage the properties of artifacts to enrich interactions and experiences with CAs, such as to engage users in co-performance with them. Likewise, to digital multi-component designs to become an engaging counterpart for humans [186, 279, 345], this thesis investigated a CA embedded into a mirror as an ecology of different resources but in line with the physical properties of the mirror and its situated use in a bathroom (cf. Chapter 6). Technology-driven communication and organizational information often take precedence over supporting the residents' alternative needs [310, 291, 5, 75].

By exploring practices and their performances with a particular focus on materials, we were able to reframe the design purpose of digitization and highlight the essential qualities of the artifact, taking into account the limitations and possibilities of the material (cf. Chapter 6). A digitally enhanced mirror, for example, can provide a space for calmness and simultaneously engaging experiences in one's current activities [268]. However, the usefulness of the mirror surface and provided applications depend on the space and context, such as a calendar would make a valuable feature on a decorative mirror in the living room instead an organizational task in a relaxed room.

The merging of traditional materializations of artifacts with technology should not strive to conceal but emphasize the meaning of practices and opportunities to contribute to new meanings and competences. Thereby the CA can either embrace further non-digital capabilities and physically interact with, for example, glancing surfaces like a mirror (cf. Chapter 6) or heat-absorbing properties like a pan (cf. Chapter 8). In this interactive cooking study, we further reflected on the impact of automation regarding the competence and traditional materials that hold meaning (cf. Chapter 10.2.2.4). In general, users get the option to interact with physical affordances of the CA, allowing to strengthen the co-performance of both actors.

By thoroughly examining domestic practices and considering the materiality of objects, we can create interactive artifacts that hold personal value for users. Acknowledging varying needs for calmness and engagement reinforces the design for performances of practice and contributes to meaningful interactions with the conversational agent. Further, this design approach requires combining automation and human agency and providing adaptive resources that emphasize meaning, materials, and competences. Against this background, CAs might be purposefully integrated into homes and enable co-performances centered around distinct artifacts and rooms to promote engaging co-experiences.

# 11 Conclusion

## 11.1 Summary of the Thesis

This research focuses on CAs and explores the design space for engaging interaction and experiences created through opportunities for human-centered co-performance. The thesis is going to address this topic in four parts:

**Part I** provides an overview of the main objectives, ideas, and methodologies explored in this thesis. In the initial Chapter 1, the thesis presents the rationale for shifting the design of voice assistants from being sheer sources of information to becoming interactive agents that co-perform with humans to create engaging and meaningful experiences. The contributions in Chapter 1.2 and the related work in Chapter 2 provide an overview of the central theories and recent research in the field of conversational user interface design, practice-based computing, and ubiquitous and personal computing. The final Chapter 3 in this part outlines the methodology used to address the research questions and describes the applied research and design activities in detail.

The following **Part II** continues to present the principal studies that contributed to this work. Chapter 5 provides the empirical grounding for the contextualization and sensitization of the field of conversational studies. Afterward, four design case studies are presented with a focus on social practice theory, conversational design, voice assistants, multimodal interaction, and co-performance. However, all studies encompass extensive prestudies, prototyping activities, and evaluations. Chapter 6 focuses on the augmentation of everyday objects with conversational agents, Chapter 7 explores a particular conversational agent for assessing food freshness, Chapter 8 introduces a comprehensive proactive and multimodal cooking assistant, and finally, Chapter 9 presents a design approach that enhances the interaction and experience with conversational agents by incorporating sound overlays.

This **Part III** concludes the thesis with a comparative discussion of the main findings of all studies and provides design implications in every chapter. Starting with the expectations and practices of users, the first Chapter 10.1 *Envisioning Conversational Agents* explores the prerequisites of voice-first interactions and becoming a conversational agent in the first place. The next Chapter 10.2 *Engaging with Conversational Agents in Co-Performance* leans on Social Practice Theory and argues conceiving conversational agents as Carriers of Practices that engage with users in co-performance to negotiate knowledge, meaning, and experience. The third Chapter 10.3 *Designing for Engaging Experiences* analyzes the advantages of multimodality to express proactive behavior and how to add modalities to enrich the voice-first experience. Lastly, this final chapter aims to provide a concise summary of the contributions presented in this work while critically examining its limitations and suggesting potential areas for future research.

## 11.2 Contributions

The objectives and research questions of this thesis, described above, are mainly associated with the research fields of Conversational User Interfaces, Practice-based Computing, and Ubiquitous and Personal Computing. Subsequently, the contributions will be categorized and summarized based on their respective areas of contributions and research question (cf. Chapter 1.2).

### 11.2.1 Conversational User Interfaces

Research in CUI focuses on transitioning from human-human conversations to interactions with conversational agents, which requires redefining norms, rules, and expectations beyond human conversations [61]. Previous work highlights the lack of design guidelines for voice-first interactions, and the challenge is further compounded by multi-componental systems like Amazon's Alexa or Google Assist. The absence or limited visual output channels pose difficulties for users in adapting to new interaction paradigms and processing information, raising the following research question: *RQ1 How might Voice Assistants become Co-performing Agents next to humans?* The contributions of this research can be categorized into two main fields. The theoretical contribution lies in envisioning conversational agents by examining human practices and utilizing the notions of informative, communicative, and expressive behaviors [17, 353]. On the other hand, the empirical contribution focuses on understanding the particularities of voice-first interactions. We offer valuable insights into users' understanding and perception of conversational agents as they handle multiple skills to provide task-oriented support in everyday household tasks. The additional empirical evaluations of our own four prototypes led to several design implications for voice-first interactions and users' envisioned skills of future conversational agents.

### 11.2.2 Practice-based Computing

This work proposes a three-step approach called Design Case Studies to create interactions through extensive empirical investigations and evaluations in specific human contexts. By building the design upon this practice-based computing approach, conversational agents might evolve into carriers of practice, offering situated learning, knowledge negotiation, and decision support. In summary, this thesis encompasses four separate design case studies to gain a comprehensive understanding of human household practices, and to answer the second research question *RQ2 How can we design for a conversational co-performance of practices?*. The theoretical contribution of this research lies in the application of the lens of Social Practice Theory to gain a deeper understanding of human practices and performances. By utilizing this theory, this work envisions the potential role of conversational agents as carriers of practices capable of facilitating engaging and meaningful interactions. Furthermore, the research emphasizes the importance of embedded negotiation processes of knowledge, meaning, and experience essential to the successful co-performance between humans and conversational agents, and, hence, the transformation of human practices [306]. The exploration of these interactions goes beyond simple single commands and sheds light on future challenges and risks associated with co-performances. Moreover, this design contribution exemplifies the practical application of the insights by designing four distinct prototypes. The ensuing comparison in the discussion of this work highlights the significance of contextual and situated research before the prototyping and implementation process, particularly considering the cognitive and social aspects of language. Furthermore, the work offers design implications to empower humans by enhancing their competence to sense, think, and act.

### 11.2.3 Ubiquitous and Personal Computing

Current research in Personal and Ubiquitous Computing primarily focuses on implementing IoT design concepts for functional purposes, such as monitoring energy consumption, enhancing home security, and controlling various aspects of smart homes. However, as the digitization of homes progresses, research and design need to investigate the home as a private space, contrasting early studies of workplaces. Moreover, home practices are grounded in enjoyable and meaningful practices that need different support through technology, leading to the last research question: *RQ3 How can multimodal agents contribute to an engaging co-performance?* Within the design of a smart kitchen assistant, this work contributes with a classification of four user groups that emphasize the need for designing for competence and autonomy. Co-performance between humans and conversational agents does not occur without risks. Hence, this thesis offers directions to balance human agency with proactive agents and smart appliances. Furthermore, by designing multimodal artifacts, the findings lead to design implications for a complementing combination of in- and output modalities. The methodological design approach to sonificate data contributes to the extension of the current CA design to complement speech and enrich the experience of information and communication. Finally, alternative visions for smart home design that contribute to proactive people are proposed.

## 11.3 Limitations and Future Work

A number of limitations of the studies should be mentioned that serve as areas for future research and design. Foremost the restrictions arising from the implementation of the prototypes and the qualitative research approach. As follows, we critically reflect on the methodology and design contributions.

### 11.3.1 Prototyping in the Wild

In summary, five comprehensive qualitative prestudies were conducted to gain a deep understanding of the design space and establish a solid foundation to develop four prototypes across four Design Case Studies. However, all prototypes differed in the implementation status as some served as Research through Design to investigate user needs and expectations. The acquired knowledge sensitized the following design case studies in Chapter 6, 7, and 8 but was not intended to modify the already existing prototypes. For example, the study in Chapter 9 investigated the impact of information sonification and complementing speech but did not yet implement the design on a smart speaker platform to run studies in the wild. Future work should investigate use cases of actual weather conditions and fitting home routines. Besides, voice modulation is a promising design case to enrich the experience of sonification. Additionally, the study in Chapter 6 proposed to integrate conversational agents in everyday objects like a smart mirror. Therefore, one alternative vision of smart homes was investigated. However, homes provide culturally rich variations of physical objects and their uses. Future prototypes might highlight further digital capabilities that the original objects inherit and provide well-being and comfort. In line with this gap, we call for more in-depth studies to extend conversational agents by multimodal interactions expanding beyond human-likeness. Further, multimodal interactions might express different levels of proactivity and need more research into the effects of complementing in- and output modalities. All of this points to the fact that this scope of research did not provide insights into the long-term effects of the prototypes or appropriation. Utilizing a Wizard-of-Oz setup contributed to an accurate understanding and response to the participants' intentions. The wizard, who observed the participants closely, interpreted their spoken requests to the assistant, creating an ideal environment for intention recognition, see Chapter 8, 7, and 6. Particularly concerning learning outcomes, transformation of practices, or agency sharing in performances, we can only speculate on the benefits and risks. However, the findings of this thesis are promising, and with the rise of LLMs [19], future work might include human-centered performances of practices that can be studied in living labs or short but targeted prototyping sessions in the wild.

### 11.3.2 Extending Generalizability

As previously mentioned, we opted for a qualitative research approach to thoroughly investigate the context and users, which enabled this research to derive strong design implications. The findings and implications of this work build a foundation for future development of robust design guidelines and principles [219, 220, 233, 61], by employing formative usability studies (cf. Chapter 7), quality experience measurements (cf. Chapter 9), and general statistical validation of interaction issues. Furthermore, the first constitutive study (cf. 5) evaluated the interaction with Amazon's Alexa. Hence, future studies should incorporate different platforms and eco-systems, as well as research on an extended set of practices. Finally, users expect personalized advice and conversations with the conversational agent, which implies studying diverse cultural backgrounds that account for varying social practices and language use [15, 344, 297]. So far, the samples were predominantly German, with a limited number of participants and international background. Chapter 9 highlighted the significance of social meaning and, in line with that, the potential appearance of misunderstandings or experiences.

Nonetheless, German, as one of many worldwide languages, shows comparable results to generalize in a predominantly English-driven research field. In summary, this qualitative approach of the thesis provides practical and theoretical implications for the future field of conversational design and co-performance.

# References

[1] Akçay, M. B., and Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication 116* (2020), 56–76.

[2] Altarriba, F., Lanzani, S. E., Torralba, A., and Funk, M. The grumpy bin: Reducing food waste through playful social interactions. In *Proceedings of the 2016 ACM Conference Companion Publication on Designing Interactive Systems - DIS '17 Companion* (2017), ACM Press, pp. 90–94.

[3] Altarriba Bertran, F., Jhaveri, S., Lutz, R., Isbister, K., and Wilde, D. Making sense of human-food interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), CHI '19, Association for Computing Machinery, p. 1–13.

[4] Amazon Europe Core S.à r.l. Amazon.de: Essen & trinken: Alexa Skills, 2019.

[5] Ambe, A. H., Brereton, M., Soro, A., Chai, M. Z., Buys, L., and Roe, P. Older People Inventing their Personal Internet of Things with the IoT Un-Kit Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 5 2019), ACM, pp. 1–15.

[6] Ammari, T., Kaye, J., Tsai, J. Y., and Bentley, F. Music, search, and iot: How people (really) use voice assistants. *ACM Trans. Comput.-Hum. Interact. 26*, 3 (apr 2019).

[7] Ardito, C., Buono, P., Costabile, M. F., and Desolda, G. Interaction with Large Displays. *ACM Computing Surveys 47*, 3 (4 2015), 1–38.

[8] Athira, S., Francis, F., Raphel, R., Sachin, N. S., Porinchu, S., and Francis, S. Smart mirror: A novel framework for interactive display. In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)* (3 2016), IEEE, pp. 1–6.

[9] Aylett, M. P., Clark, L., and Cowan, B. R. Siri, echo and performance: You have to suffer darling. In *Conf. Hum. Factors Comput. Syst. - Proc.* (New York, New York, USA, 2019), ACM Press, pp. 1–10.

[10] Aylett, M. P., Vazquez-Alvarez, Y., Kristensson, P. O., and Whittaker, S. None of a CHInd: Relationship counselling for HCI and speech technology. In *Conf. Hum. Factors Comput. Syst. - Proc.* (2014), pp. 749–758.

[11] Bader, G. E., and Rossi, C. A. *Focus groups: A step-by-step guide*. Bader Group, 1998.

[12] Barko-Sherif, S., Elsweiler, D., and Harvey, M. Conversational agents for recipe recommendation. *CHIIR 2020 - Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (2020), 73–82.

[13] Barton, K. L., Wrieden, W. L., and Anderson, A. S. Validity and reliability of a short questionnaire for assessing the impact of cooking skills interventions. *Journal of Human Nutrition and Dietetics 24*, 6 (2011), 588–595.

[14] Baurley, S., Petreca, B., Selinas, P., Selby, M., and Flintham, M. Modalities of expression: Capturing embodied knowledge in cooking. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction* (New York, NY, USA, 2020), TEI'20, ACM, pp. 785–797.

[15] Bavelas, J., Gerwing, J., Allison, M., and Sutton, C. Dyadic evidence for grounding with abstract deictic gestures. *Integrating gestures: The interdisciplinary nature of gestures* (2011).

[16] Bavelas, J., Gerwing, J., and Healing, S. Doing mutual understanding. calibrating with micro-sequences in face-to-face dialogue. *Journal of Pragmatics 121* (2017), 91–112.

[17] Bavelas, J. B. Behaving and communicating: A reply to motley. *Western Journal of Speech Communication 54*, 4 (1990), 593–602.

[18] Becker, H. S. Becoming a marihuana user. *American journal of Sociology 59*, 3 (1953), 235–242.

[19] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021), FAccT '21, Association for Computing Machinery, p. 610–623.

[20] Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., and Lottridge, D. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol. 2*, 3 (2018), 1–24.

[21] Berkholz, J., Esau-Held, M., Boden, A., Stevens, G., and Tolmie, P. Becoming an online wine taster: An ethnographic study on the digital mediation of taste. *Proc. ACM Hum.-Comput. Interact. 7*, CSCW1 (apr 2023).

[22] Berkholz, J., Esau-Held, M., and Stevens, G. Negotiating taste for digital depiction: Aligning individual concepts of taste perception in a co-design process. In *Proceedings of Mensch Und Computer 2022* (New York, NY, USA, 2022), MuC '22, Association for Computing Machinery, p. 137–146.

[23] Biernacki, P., and Waldorf, D. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research 10*, 2 (1981), 141–163.

[24] Biggs, L., Juravle, G., and Spence, C. Haptic exploration of plateware alters the perceived texture and taste of food. *Food Quality and Preference 50* (2016), 129–134.

[25] BILLINGHURST, M. Put That Where? Voice and Gesture at the Graphics Interface. *Comput. Graph. 32*, 4 (nov 1998), 60–63.

[26] BIRDWHISTELL, R. L. Contribution of linguistic-kinesic studies to the understanding of schizophrenia. *Schizophrenia: An integrated approach.* (1959).

[27] BLATTNER, M., SUMIKAWA, D., AND GREENBERG, R. Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interact. 4*, 1 (mar 1989), 11–44.

[28] BOLGER, N., DAVIS, A., AND RAFAELI, E. Diary Methods: Capturing Life as it is Lived. *Annual Review of Psychology 54*, 1 (2 2003), 579–616.

[29] BORGIA, E. The internet of things vision: Key features, applications and open issues. *Computer Communications 54* (2014), 1–31.

[30] BOURDIEU, P. *Distinction: A social critique of the judgement of taste*. Harvard university press, Cambridge, MA, USA, 1984.

[31] BOWERS, J. The logic of annotated portfolios: Communicating the value of 'research through design'. In *Proceedings of the Designing Interactive Systems Conference, DIS '12* (New York, New York, USA, 2012), ACM Press, pp. 68–77.

[32] BRAUN, V., AND CLARKE, V. Using thematic analysis in psychology. *Qualitative Research in Psychology 3*, 2 (1 2006), 77–101.

[33] BREAZEAL, C. Emotion and sociable humanoid robots. *International journal of human-computer studies 59*, 1-2 (2003), 119–155.

[34] BREWSTER, S. A. Providing a structured method for integrating non-speech audio into human-computer interfaces.

[35] BRUCE, A., NOURBAKHSH, I., AND SIMMONS, R. The role of expressiveness and attention in human-robot interaction. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)* (2002), vol. 4, pp. 4138–4142 vol.4.

[36] BSH HAUSGERäTE GMBH. Häufige Fragen zum Cookit, 2021.

[37] BURKHARDT, F., AND STEGMANN, J. Emotional speech synthesis: Applications, history and possible future. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2009* (2009), 190–199.

[38] BURMESTER, M., ZEINER, K., SCHIPPERT, K., AND PLATZ, A. Creating positive experiences with digital companions. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), CHI EA '19, Association for Computing Machinery, p. 1–6.

[39] CABRAL, J. P., COWAN, B. R., ZIBREK, K., AND MCDONNELL, R. The Influence of Synthetic Voice on the Evaluation of a Virtual Character. In *Proc. Interspeech 2017* (2017), pp. 229–233.

[40] CABRAL, J. P., AND REMIJN, G. B. Auditory icons: Design and physical characteristics. *Appl. Ergon. 78*, January (jul 2019), 224–239.

[41] CAMPBELL, I. G. Basal emotional patterns expressible in music. *The American Journal of Psychology 55*, 1 (1942), 1–17.

[42] CANDELLO, H., MUNTEANU, C., CLARK, L., SIN, J., TORRES, M. I., PORCHERON, M., MYERS, C. M., COWAN, B., FISCHER, J., SCHLÖGL, S., MURAD, C., AND REEVES, S. CUI@CHI: Mapping grand challenges for the conversational user interface community. In *Conf. Hum. Factors Comput. Syst. - Proc.* (New York, NY, USA, apr 2020), ACM, pp. 1–8.

[43] CANDELLO, H., PINHANEZ, C., AND FIGUEIREDO, F. Typefaces and the perception of humanness in natural language chatbots. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI '17, Association for Computing Machinery, p. 3476–3487.

[44] CARDELLO, A. V., AND SCHUTZ, H. G. The concept of food freshness: Uncovering its meaning and importance to consumers. In *ACS Symposium Series*, vol. 836. oct 2002, pp. 22–41.

[45] CARNAP, R. *Introduction to semantics and formalization of logic*. Harvard University Press, 1959.

[46] CARVALHO, F. R., STEENHAUT, K., VAN EE, R., TOUHAFI, A., AND VELASCO, C. Sound-enhanced gustatory experiences and technology. In *Proc. 1st Work. Multi-sensorial Approaches to Human-Food Interact.* (New York, NY, USA, nov 2016), {MHFI} '16, ACM, pp. 1–8.

[47] CASTELLI, N., OGONOWSKI, C., JAKOBI, T., STEIN, M., STEVENS, G., AND WULF, V. What Happened in my Home? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 5 2017), vol. 2017-May, ACM, pp. 853–866.

[48] CEDILLO, P., SANCHEZ, C., CAMPOS, K., AND BERMEO, A. A Systematic Literature Review on Devices and Systems for Ambient Assisted Living: Solutions and Trends from Different User Perspectives. In *2018 International Conference on eDemocracy & eGovernment (ICEDEG)* (4 2018), IEEE, pp. 59–66.

[49] CHAVEZ-SANCHEZ, F., FRANCO, G. A. M., DE LA PEÑA, G. A. M., AND CARRILLO, E. I. H. Beyond What is Said. In *Proc. 2nd Conf. Conversational User Interfaces* (New York, NY, USA, jul 2020), ACM, pp. 1–3.

[50] CHEFKOCH GMBH. Die Chefkoch-App für Smartphone und Tablet, 2022.

[51] CHEN, Y., MAO, Z., AND QIU, J. L. *Super-sticky WeChat and Chinese society*, 1 ed. Emerald Publishing Limited, Howard House, Wagon Lane, Bingley BD16 1WA, UK, 2018.

[52] CHION, M. *Audio-vision: Sound on Screen*. Columbia University Press, News York, USA, 1994.

[53] CHO, M., LEE, S.-S., AND LEE, K.-P. Once a kind friend is now a thing: Understanding how conversational agents at home are forgotten. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (New York, NY, USA, 2019), DIS '19, Association for Computing Machinery, p. 1557–1569.

[54] CHO, M., AND SAAKES, D. Calm automaton: A DIY toolkit for ambient displays. *Conference on Human Factors in Computing Systems - Proceedings Part F1276*, c (2017), 393–396.

[55] CHUNG, C.-F., AGAPIE, E., SCHROEDER, J., MISHRA, S., FOGARTY, J., AND MUNSON, S. A. When personal tracking becomes social: Examining the use of instagram for healthy eating. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI'17, ACM, pp. 1674–1687.

[56] CHUNG, D., TSAI, W.-C., LIANG, R.-H., KONG, B., HUANG, Y., CHANG, F.-C., AND LIU, M. Designing Auditory Experiences for Technology Imagination. In *32nd Aust. Conf. Human-Computer Interact.* (New York, NY, USA, dec 2020), ACM, pp. 682–686.

[57] CILA, N., SMIT, I., GIACCARDI, E., AND KRÖSE, B. Products as Agents. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 5 2017), vol. 2017-May, ACM, pp. 448–459.

[58] CLARK, H., BRENNAN, S., RESNICK, L., LEVINE, J., AND TEASLEY, S. Grounding in communication perspectives on socially shared cognition (pp. 127–149). *Washington, DC, US: American Psychological Association 35* (1991).

[59] CLARK, H. H. *Using language*. Cambridge university press, 1996.

[60] CLARK, L., DOYLE, P., GARAIALDE, D., GILMARTIN, E., SCHLÖGL, S., EDLUND, J., AYLETT, M., CABRAL, J., MUNTEANU, C., EDWARDS, J., AND R COWAN, B. The State of Speech in HCI: Trends, Themes and Challenges. *Interact. Comput. 31*, 4 (dec 2019), 349–371.

[61] CLARK, L., MUNTEANU, C., WADE, V., COWAN, B. R., PANTIDI, N., COONEY, O., DOYLE, P., GARAIALDE, D., EDWARDS, J., SPILLANE, B., GILMARTIN, E., AND MURAD, C. What Makes a Good Conversation? In *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst. - CHI '19* (New York, New York, USA, 2019), ACM Press, pp. 1–12.

[62] CLARKE, V., BRAUN, V., AND HAYFIELD, N. Thematic analysis. *Qualitative psychology: A practical guide to research methods* (2015), 222–248.

[63] COLANTONIO, S., COPPINI, G., GERMANESE, D., GIORGI, D., MAGRINI, M., MARRACCINI, P., MARTINELLI, M., MORALES, M. A., PASCALI, M. A., RACCICHINI, G., RIGHI, M., AND SALVETTI, O. A smart mirror to promote a healthy lifestyle. *Biosystems Engineering 138* (10 2015), 33–43.

[64] COMBER, R., HOONHOUT, J., VAN HALTEREN, A., MOYNIHAN, P., AND OLIVIER, P. Food practices as situated action: Exploring and designing for everyday food practices with households. CHI '13, Association for Computing Machinery, p. 2457–2466.

[65] COMBER, R., THIEME, A., RAFIEV, A., TAYLOR, N., KRäMER, N., AND OLIVIER, P. Bincam: Designing for engagement with facebook for behavior change. In *IFIP Conference on Human-Computer Interaction* (2013), Springer, pp. 99–115.

[66] CORBETT, E., AND WEBER, A. What can I say? In *Proc. 18th Int. Conf. Human-Computer Interact. with Mob. Devices Serv. - MobileHCI '16* (New York, New York, USA, 2016), ACM Press, pp. 72–82.

[67] COSTELL, E. A comparison of sensory methods in quality control. *Food Quality and Preference 13*, 6 (2002), 341–353.

[68] COWAN, B. R., PANTIDI, N., COYLE, D., MORRISSEY, K., CLARKE, P., AL-SHEHRI, S., EARLEY, D., AND BANDEIRA, N. "What can i help you with?": Infrequent users' experiences of intelligent personal assistants. In *Proc. 19th Int. Conf. Human-Computer Interact. with Mob. Devices Serv. MobileHCI 2017* (New York, New York, USA, 2017), ACM Press, pp. 1–12.

[69] CRABTREE, A., AND RODDEN, T. Domestic Routines and Design for the Home. *Computer Supported Cooperative Work (CSCW) 13*, 2 (4 2004), 191–220.

[70] DAFEI, D. An exploration of the relationship between learner autonomy and english proficiency. *Asian EFL Journal 24*, 4 (2007), 24–34.

[71] DAVID, B., CHALON, R., ZHANG, B., AND YIN, C. Design of a collaborative learning environment integrating emotions and virtual assistants (chatbots). In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (2019), pp. 51–56.

[72] DE ANGELI, A., PETRELLI, D., ET AL. Bridging the gap between nlp and hci: A new synergy in the name of the user. In *Proceedings of the CHI 2000 Workshop on Natural Language Interfaces* (2000), vol. 4.

[73] DEMAEGHT, A., NERB, J., AND MüLLER, A. A survey-based study to identify user annoyances of german voice assistant users. In *International Conference on Human-Computer Interaction* (Cham, 2022), Springer, p. 261–271.

[74] DERESHEV, D., KIRK, D., MATSUMURA, K., AND MAEDA, T. Long-term value of social robots through the eyes of expert users. In *Proceedings of the 2019 CHI Conference on*

*Human Factors in Computing Systems - CHI '19* (New York, New York, USA, 2019), no. Chi, ACM Press, pp. 1–12.

[75] DESJARDINS, A., VINY, J. E., KEY, C., AND JOHNSTON, N. Alternative avenues for IoT: Designing with non-stereotypical homes. In *Conference on Human Factors in Computing Systems - Proceedings* (New York, New York, USA, 2019), ACM Press, pp. 1–13.

[76] DESJARDINS, A., WAKKARY, R., AND ODOM, W. Investigating Genres and Perspectives in HCI Research on the Home. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 4 2015), vol. 2015-April, ACM, pp. 3073–3082.

[77] DIEFENBACH, S., AND HASSENZAHL, M. Werkzeuge für Gestaltung und Evaluation auf der Erlebnisebene. In *Psychologie in der nutzerzentrierten Produktgestaltung*. Springer, Berlin, 2017, pp. 157–169.

[78] DOLEJŠOVá, M., BERTRAN, F. A., WILDE, D., AND DAVIS, H. Crafting and tasting issues in everyday human-food interactions. In *DIS'19 Companion: Companion Publication of the 2019 ACM Designing Interactive Systems Conference (San Diego, California, USA)* (New York, NY, USA, 2019), ACM, pp. 361–364.

[79] DOLEJŠOVá, M., WILDE, D., ALTARRIBA BERTRAN, F., AND DAVIS, H. Disrupting (more-than-) human-food interaction: Experimental design, tangibles and food-tech futures. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (New York, NY, USA, 2020), DIS '20, Association for Computing Machinery, p. 993–1004.

[80] DONALD, R., KREUTZ, G., MITCHELL, L., AND MACDONALD, R. What is music health and wellbeing and why is it important? In *Music, Health, and Wellbeing*. Oxford University Press, Oxford, 2012, pp. 3–11.

[81] DÖRRENBäCHER, J., LÖFFLER, D., AND HASSENZAHL, M. Becoming a robot-overcoming anthropomorphism with techno-mimesis. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (2020), pp. 1–12.

[82] DOW, S., SAPONAS, T. S., LI, Y., AND LANDAY, J. A. External representations in ubiquitous computing design and the implications for design tools. In *Proceedings of the 6th Conference on Designing Interactive Systems* (New York, NY, USA, 2006), DIS'06, ACM, p. 241–250.

[83] DRUGA, S., BREAZEAL, C., WILLIAMS, R., AND RESNICK, M. "Hey Google is it ok if I eat you?" Initial explorations in child-agent interaction. In *IDC 2017 - Proc. 2017 ACM Conf. Interact. Des. Child.* (New York, NY, USA, jun 2017), ACM, pp. 595–600.

[84] DUGUID, P. "the art of knowing": Social and tacit dimensions of knowledge and the limits of the community of practice. *The Information Society 21* (2005), 109 – 118.

[85] ECCLES, D. W., AND ARSAL, G. The think aloud method: what is it and how do i use it? *Qualitative Research in Sport, Exercise and Health 9*, 4 (2017), 514–531.

[86] EGGEN, B., VAN DEN HOVEN, E., AND TERKEN, J. Human-Centered Design and Smart Homes: How to Study and Design for the Home Experience? In *Handbook of Smart Homes, Health Care and Well-Being*. Springer International Publishing, Cham, 2014, pp. 1–9.

[87] EIDE, E., AARON, A., BAKIS, R., HAMZA, W., PICHENY, M., AND PITRELLI, J. A corpus-based approach to expressive speech synthesis. In *Fifth ISCA Workshop on Speech Synthesis* (2004).

[88] ENGE, K., RIND, A., IBER, M., HÖLDRICH, R., AND AIGNER, W. It's about Time: Adopting Theoretical Constructs from Visualization for Sonification. In *Audio Most. 2021* (New York, NY, USA, sep 2021), ACM, pp. 64–71.

[89] ESAU, M., KRAUSS, V., LAWO, D., AND STEVENS, G. Losing its touch: Understanding user perception of multimodal interaction and smart assistance. In *Designing Interactive Systems Conference* (New York, NY, USA, 2022), DIS '22, ACM, pp. 1288—-1299.

[90] ESAU, M., LAWO, D., CASTELLI, N., JAKOBI, T., AND STEVENS, G. Morning Routines between Calm and Engaging: Designing a Smart Mirror. In *Proceedings of the 5th International Conference on Computer-Human Interaction Research and Applications* (Setúbal, Portugal, 2021), SCITEPRESS - Science and Technology Publications, pp. 58–69.

[91] ESAU, M., LAWO, D., NEIFER, T., STEVENS, G., AND BODEN, A. Trust your guts: fostering embodied knowledge and sustainable practices through voice interaction. *Personal and Ubiquitous Computing 27* (2022), 415–434.

[92] ESAU, M., LAWO, D., AND STEVENS, G. Really smart fridges: Investigating sustainable household storage practices. In *ICT4S Poster Session* (2020).

[93] FARR-WHARTON, G., CHOI, J. H.-J., AND FOTH, M. Technicolouring the fridge: Reducing food waste through uses of colour-coding and cameras. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia - MUM '14* (Melbourne, Victoria, Australia, 2014), ACM Press, pp. 48–57.

[94] FISCHER, J. E., REEVES, S., PORCHERON, M., AND SIKVELAND, R. O. Progressivity for voice interface design. In *Proc. 1st Int. Conf. Conversational User Interfaces - CUI '19* (New York, New York, USA, 2019), no. ii, ACM Press, pp. 1–8.

[95] FORLIZZI, J., AND BATTARBEE, K. Understanding experience in interactive systems. In *Proc. 5th Conf. Des. Interact. Syst. Process. Pract. methods, Tech.* (New York, NY, USA, aug 2004), ACM, pp. 261–268.

[96] FRANKE, N., AND SHAH, S. How communities support innovative activities: an exploration of assistance and sharing among end-users. *Research Policy 32*, 1 (2003), 157–178.

[97] FREDERKING, R. E. Grice's maxims: do the right thing. *Frederking, RE* (1996).

[98] FUCHSBERGER, V., MURER, M., AND TSCHELIGI, M. Materials, materiality, and media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 4 2013), ACM, pp. 2853–2862.

[99] FUENTES, C., PORCHERON, M., FISCHER, J. E., COSTANZA, E., MALILK, O., AND RAMCHURN, S. D. Tracking the consumption of home essentials. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk, 2019), ACM Press, pp. 1–13.

[100] FUJINAMI, K., KAWSAR, F., AND NAKAJIMA, T. AwareMirror: A Personalized Display Using a Mirror. In *Lecture Notes in Computer Science*, vol. 3468 of *PERVASIVE'05*. Springer-Verlag, Berlin, Heidelberg, 2005, pp. 315–332.

[101] GANGLBAUER, E., FITZPATRICK, G., AND COMBER, R. Negotiating food waste: Using a practice lens to inform design. *ACM Trans. Comput.-Hum. Interact. 20*, 2 (May 2013).

[102] GARG, R., AND SENGUPTA, S. Conversational technologies for in-home learning: Using co-design to understand children's and parents' perspectives. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, apr 2020), ACM, pp. 1–13.

[103] GATTI, E., AND RICHTER, C. *WeChat – Die chinesische Super-App*. Springer Fachmedien Wiesbaden, Wiesbaden, 2019, pp. 23–30.

[104] GAVER, B., AND MARTIN, H. Alternatives: Exploring information appliances through conceptual design proposals. *Conference on Human Factors in Computing Systems - Proceedings 2*, 1 (2000), 209–216.

[105] GAVER, W. Designing for Ludic Aspects of Everyday Life. *ERCIM New 47* (2001), 20–21.

[106] GAVER, W. What should we expect from research through design? *Conference on Human Factors in Computing Systems - Proceedings* (2012), 937–946.

[107] GAVER, W. W. The SonicFinder: An Interface That Uses Auditory Icons. *Human-Computer Interact. 4*, 1 (mar 1989), 67–94.

[108] GHERARDI, S. Situated knowledge and situated action: What do practice-based studies promise. *The SAGE handbook of new approaches in management and organization* (2008), 516–525.

[109] GHERARDI, S., AND NICOLINI, D. To transfer is to transform: The circulation of safety knowledge. *Organization 7*, 2 (2000), 329–348.

[110] GIACCARDI, E., AND FISCHER, G. Creativity and evolution: a metadesign perspective. *Digital Creativity 19*, 1 (2008), 19–32.

[111] GIACCARDI, E., SPEED, C., CILA, N., AND CALDWELL, M. Things as co-ethnographers: Implications of a thing perspective for design and anthropology. *Design anthropological futures 235* (2016).

[112] GIDDENS, A., AND GIDDENS, A. Agency, structure. *Central Problems in Social Theory: Action, structure and contradiction in social analysis* (1979), 49–95.

[113] GNEWUCH, U., MORANA, S., AND MAEDCHE, A. Towards designing cooperative and social conversational agents for customer service. In *Proceedings of the International Conference on Information Systems (ICIS) 2017* (Seoul, South Korea, 2017), vol. 6, Association for Information Systems (AIS), pp. 4046–4058.

[114] GRAESSER, A. C., DOWELL, N., AND CLEWLEY, D. Assessing collaborative problem solving through conversational agents. In *Innovative assessment of collaboration*. Springer, 2017, pp. 65–80.

[115] GRAM-HANSSEN, K. Understanding change and continuity in residential energy consumption. *Journal of Consumer Culture 11*, 1 (mar 2011), 61–78.

[116] GREEN, P., AND WEI-HAAS, L. The rapid development of user interfaces: Experience with the wizard of oz method. In *Proceedings of the Human Factors Society Annual Meeting* (1985), vol. 29, SAGE Publications Sage CA: Los Angeles, CA, pp. 470–474.

[117] GREEN, W., GYI, D., KALAWSKY, R., AND ATKINS, D. Capturing user requirements for an integrated home environment. *ACM International Conference Proceeding Series 82* (2004), 255–258.

[118] GRIMES, A., AND HARPER, R. Celebratory technology: New directions for food research in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2008), CHI '08, ACM, pp. 467–476.

[119] GRUDIN, J., AND JACQUES, R. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), Association for Computing Machinery, p. 1–11.

[120] GRUDIN, J., AND JACQUES, R. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst. - CHI '19* (New York, New York, USA, 2019), ACM Press, pp. 1–11.

[121] Gu, H., and Wang, D. A content-aware fridge based on RFID in smart home for home-healthcare. In *2009 11th International Conference on Advanced Communication Technology* (New York, NY, USA, 2009), vol. 2, IEEE, pp. 987–990.

[122] Guest, G., MacQueen, K. M., and Namey, E. E. *Applied Thematic Analysis*. SAGE Publications, Los Angeles, USA, Nov 2011.

[123] Hallnäs, L., and Redström, J. *Interaction design: foundations, experiments*. Textile Research Centre, Swedish School of Textiles, Unversity College of~..., 2006.

[124] Hamada, R., Okabe, J., Ide, I., Satoh, S., Sakai, S., and Tanaka, H. Cooking Navi: assistant for daily cooking in kitchen. In *Proceedings of the 13th Annual ACM International Conference on Multimedia* (New York, NY, USA, 2005), MM05, ACM, pp. 371–374.

[125] Hämäläinen, P. Interactive video mirrors for sports training. In *ACM International Conference Proceeding Series* (New York, New York, USA, 2004), vol. 82, ACM Press, pp. 199–202.

[126] Hargreaves, T., and Wilson, C. Introduction: Smart Homes and Their Users. Springer, 2017, pp. 1–14.

[127] Harrar, V., and Spence, C. The taste of cutlery: how the taste of food is affected by the weight, size, shape, and colour of the cutlery used to eat it. *Flavour 2*, 1 (2013), 1–13.

[128] Hassenzahl, M. User experience (ux) towards an experiential perspective on product quality. In *Proceedings of the 20th Conference on l'Interaction Homme-Machine* (New York, NY, USA, 2008), IHM'08, ACM, pp. 11–15.

[129] Hassenzahl, M. Experience Design: Technology for All the Right Reasons. *Synth. Lect. Human-Centered Informatics 3*, 1 (jan 2010), 1–95.

[130] Hassenzahl, M., Borchers, J., Boll, S., der Pütten, A. R.-v., and Wulf, V. Otherware: how to best interact with autonomous systems. *Interactions 28*, 1 (2021), 54–57.

[131] Hassenzahl, M., Diefenbach, S., and Göritz, A. Needs, affect, and interactive products - Facets of user experience. *Interact. Comput. 22*, 5 (2010), 353–362.

[132] Hassenzahl, M., Eckoldt, K., Diefenbach, S., Laschke, M., Lenz, E., and Kim, J. Designing moments of meaning and pleasure. Experience design and happiness. *International Journal of Design 7*, 3 (2013), 21–31.

[133] Hassenzahl, M., and Klapperich, H. Convenient, clean, and efficient? the experiential costs of everyday automation. In *Proceedings of the 8th nordic conference on human-computer interaction: Fun, fast, foundational* (2014), pp. 21–30.

[134] HEBROK, M., AND HEIDENSTRØM, N. Contextualising food waste prevention - 'decisive' moments within everyday practices. *Journal of Cleaner Production 210* (2019-02), 1435–1448.

[135] HEDIN, B., KATZEFF, C., ERIKSSON, E., AND PARGMAN, D. A systematic review of digital behaviour change interventions for more sustainable food consumption. *Sustainability 11*, 9 (2019-05-08), 2638.

[136] HEKTNER, J. M., SCHMIDT, J. A., AND CSIKSZENTMIHALYI, M. *Experience sampling method: Measuring the quality of everyday life*. Sage, 2007.

[137] HOBERT, S., AND MEYER VON WOLFF, R. Say hello to your new automated tutor – a structured literature review on pedagogical conversational agents. *Wirtschaftsinformatik* (2019), 301–314.

[138] HOME CONNECT GMBH. Home Connect, 2021.

[139] HONER, A. Life-world analysis in ethnography. *Qualitative research* (2004), 113–117.

[140] HONG, J., YI, H. B., PYUN, J., AND LEE, W. SoundWear: Effect of non-speech sound augmentation on the outdoor play experience of children. In *DIS 2020 - Proc. 2020 ACM Des. Interact. Syst. Conf.* (New York, NY, USA, jul 2020), ACM, pp. 2201–2213.

[141] HONIG, S., AND ORON-GILAD, T. Comparing laboratory user studies and video-enhanced web surveys for eliciting user gestures in human-robot interactions. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, 2020), HRI '20, Association for Computing Machinery, p. 248–250.

[142] IACUCCI, G., KUUTTI, K., AND RANTA, M. On the move with a magic thing: Role playing in concept design of mobile services and devices. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (New York, NY, USA, 2000), DIS '00, Association for Computing Machinery, p. 193–202.

[143] ISO NORM. Acoustics — soundscape — part 1: Definition and conceptual framework, 2022. Last accessed 31.03.2022.

[144] JACOBS, R., SCHNÄDELBACH, H., JÄGER, N., LEAL, S., SHACKFORD, R., BENFORD, S., AND PATEL, R. The Performative Mirror Space. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 5 2019), ACM, pp. 1–14.

[145] JACQUES, R., FÖLSTAD, A., GERBER, E., GRUDIN, J., LUGER, E., MONROY-HERNÁNDEZ, A., AND WANG, D. Conversational agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, may 2019), ACM, pp. 1–8.

[146] JAIN, M., KUMAR, P., KOTA, R., AND PATEL, S. N. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (New York, NY, USA, 2018), DIS '18, Association for Computing Machinery, p. 895–906.

[147] JAKOBI, T., OGONOWSKI, C., CASTELLI, N., STEVENS, G., AND WULF, V. The catch(es) with smart home - Experiences of a Living Lab field study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI'17, ACM, pp. 1620–1633.

[148] JAKOBI, T., STEVENS, G., CASTELLI, N., OGONOWSKI, C., SCHAUB, F., VINDICE, N., RANDALL, D., TOLMIE, P., AND WULF, V. Evolving Needs in IoT Control and Accountability. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2*, 4 (12 2018), 1–28.

[149] JENSEN, R. H., STRENGERS, Y., RAPTIS, D., NICHOLLS, L., KJELDSKOV, J., AND SKOV, M. B. Exploring Hygge as a Desirable Design Vision for the Sustainable Smart Home. In *Proceedings of the 2018 Designing Interactive Systems Conference* (New York, NY, USA, 6 2018), ACM, pp. 355–360.

[150] JERALD, J. *The VR book: Human-centered design for virtual reality*. Morgan & Claypool, New York, USA, 2015.

[151] JUNG, H., STOLTERMAN, E., RYAN, W., THOMPSON, T., AND SIEGEL, M. Toward a framework for ecologies of artifacts. In *Proceedings of the 5th Nordic conference on Human-computer interaction building bridges - NordiCHI '08* (New York, New York, USA, 2008), vol. 358, ACM Press, p. 201.

[152] JUSLIN, P. N. What does music express? Basic emotions and beyond. *Front. Psychol. 4*, SEP (2013).

[153] JUSLIN, P. N., AND LAUKKA, P. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin 129*, 5 (2003), 770.

[154] JUSLIN, P. N., AND SLOBODA, J. *Handbook of music and emotion: Theory, research, applications*. Oxford University Press, Oxford, 2011.

[155] KALLIO, H., PIETILä, A.-M., JOHNSON, M., AND KANGASNIEMI, M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of advanced nursing 72*, 12 (2016), 2954–2965.

[156] KARRAY, F., ALEMZADEH, M., ABOU SALEH, J., AND ARAB, M. N. Human-computer interaction: Overview on state of the art. *International Journal on Smart Sensing and Intelligent Systems 1*, 1 (2008), 137–159.

[157] **Kato, F., and Hasegawa, S.** Interactive cooking simulator: Showing food ingredients appearance changes in frying pan cooking. In *Proceedings of the 5th International Workshop on Multimedia for Cooking & Eating Activities* (New York, NY, USA, 2013), CEA '13, ACM, pp. 33–38.

[158] **Kelle, U., and Erzberger, C.** *Qualitative and quantitative methods: not in opposition*. SAGE Publications, London, 2004, p. 172–177. publisher: Sage Publications London.

[159] **Kerruish, E.** Arranging sensations: smell and taste in augmented and virtual reality. *The Senses and Society 14*, 1 (2019), 31–45.

[160] **Khot, R. A., Mueller, F., et al.** Human-food interaction. *Foundations and Trends in Human–Computer Interaction 12*, 4 (2019), 238–415.

[161] **Kiesler, S., and Goetz, J.** Mental models of robotic assistants. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2002), CHI EA '02, Association for Computing Machinery, p. 576–577.

[162] **Kilian, K., and Kreutzer, R. T.** *Voice-Marketing*. Springer Fachmedien, Wiesbaden, 2022, p. 279–312.

[163] **Kim, D. J., and Lim, Y. K.** Co-performing agent: Design for building user–agent partnership in learning and adaptive services. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–14.

[164] **Kim, K., Boelling, L., Haesler, S., Bailenson, J., Bruder, G., and Welch, G. F.** Does a digital assistant need a body? the influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (New York, NY, USA, 2018), IEEE, pp. 105–114.

[165] **Kindermann, A.** Thermomix: Ein küchenklassiker wird digital. In *Fallstudien zur Digitalen Transformation: Case Studies für die Lehre und praktische Anwendung*, C. Gärtner and C. Heinrich, Eds. Springer Fachmedien, Wiesbaden, 2018, pp. 107–128.

[166] **Kingaby, S. A.** *Data-Driven Alexa Skills: Voice Access to Rich Data Sources for Enterprise Applications*, 1 ed. Springer, La Vergne, TN, USA, 2021.

[167] **Kitchen Stories.** Kitchen Stories, 2022.

[168] **Klein, A. M., Hinderks, A., Schrepp, M., and Thomaschewski, J.** Measuring user experience quality of voice assistants. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)* (Danvers, MA, 2020), IEEE, p. 1–4.

[169] **Koolagudi, S. G., and Rao, K. S.** Emotion recognition from speech: a review. *International journal of speech technology 15*, 2 (2012), 99–117.

[170] KRAMPEN, G. *Psychologie der Kreativität*, vol. 44 of *Heidelberger Jahrbücher*. Hogrefe, Berlin, Heidelberg, 2019.

[171] KRANZ, M., HOLLEIS, P., AND SCHMIDT, A. Embedded Interaction: Interacting with the Internet of Things. *IEEE Internet Computing 14*, 2 (3 2010), 46–53.

[172] KRAUS, M., SCHILLER, M., BEHNKE, G., BERCHER, P., DORNA, M., DAMBIER, M., GLIMM, B., BIUNDO, S., AND MINKER, W. "Was that successful?" on integrating proactive meta-dialogue in a diy-assistant using multimodal cues. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (New York, NY, USA, 2020), ICMI '20, ACM, pp. 585–594.

[173] KRAUS, M., WAGNER, N., CALLEJAS, Z., AND MINKER, W. The role of trust in proactive conversational assistants. *IEEE Access 9* (2021), 112821–112836.

[174] KRAUSS, V., JASCHE, F., SASSMANNSHAUSEN, S. M., LUDWIG, T., AND BODEN, A. Research and practice recommendations for mixed reality design – different perspectives from the community. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology* (New York, NY, USA, 2021), VRST '21, Association for Computing Machinery.

[175] KRAUSS, V., NEBELING, M., JASCHE, F., AND BODEN, A. Elements of xr prototyping: Characterizing the role and use of prototypes in augmented and virtual reality design. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA, 2022), CHI '22, ACM.

[176] KRYGIER, J. B. Sound and geographic visualization. In *Mod. Cartogr. Ser.*, vol. 2. Elsevier Science Ltd, Kidlington, Oxford, OX5 1GB, U.K., 1994, pp. 149–166.

[177] KUIJER, L., AND GIACCARDI, E. Co-performance: Conceptualizing the role of artificial agency in the design of everyday life. In *Conference on Human Factors in Computing Systems - Proceedings* (New York, New York, USA, 2018), vol. 2018-April, ACM Press, pp. 1–13.

[178] KUNO, Y., SADAZUKA, K., KAWASHIMA, M., YAMAZAKI, K., YAMAZAKI, A., AND KUZUOKA, H. Museum guide robot based on sociological interaction analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2007), CHI '07, Association for Computing Machinery, p. 1191–1194.

[179] KUZMINYKH, A., SUN, J., GOVINDARAJU, N., AVERY, J., AND LANK, E. Genie in the Bottle: Anthropomorphized Perceptions of Conversational Agents. In *Conf. Hum. Factors Comput. Syst. - Proc.* (New York, NY, USA, apr 2020), ACM, pp. 1–13.

[180] LAESTADIUS, L., BISHOP, A., GONZALEZ, M., ILLENČÍK, D., AND CAMPOS-CASTILLO, C. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society* (2022), 14614448221142007.

[181] LANZ, P. The concept of intelligence in psychology and philosophy. In *Prerational Intelligence: Adaptive Behavior and Intelligent Systems Without Symbols and Logic, Volume 1, Volume 2 Prerational Intelligence: Interdisciplinary Perspectives on the Behavior of Natural and Artificial Systems, Volume 3*. Springer, 2000, pp. 19–30.

[182] LASCHKE, M., NEUHAUS, R., DÖRRENBÄCHER, J., HASSENZAHL, M., WULF, V., ROSENTHAL-VON DER PÜTTEN, A., BORCHERS, J., AND BOLL, S. Otherware needs otherness: Understanding and designing artificial counterparts. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (New York, NY, USA, 2020), NordiCHI '20, ACM.

[183] LAVE, J., AND WENGER, E. *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.

[184] LAWO, D., ENGELBUTZEDER, P., ESAU, M., AND STEVENS, G. Networks of practices: Exploring design opportunities for interconnected practices. In *Proceedings of 18th European Conference on Computer-Supported Cooperative Work* (2020), European Society for Socially Embedded Technologies (EUSSET).

[185] LAWO, D., ESAU, M., ENGELBUTZEDER, P., AND STEVENS, G. Going Vegan: The Role(s) of ICT in Vegan Practice Transformation. *Sustainability 12*, 12 (jun 2020), 5184.

[186] LAZARO, M. J., KIM, S., LEE, J., CHUN, J., KIM, G., YANG, E., BILYALOVA, A., AND YUN, M. H. A review of multimodal interaction in intelligent systems. In *Human-Computer Interaction. Theory, Methods and Tools* (Berlin, Heidelberg, 2021), vol. 12762 of *Lecture Notes in Computer Science*, Springer, pp. 206–219.

[187] LEE, C.-C., KIM, J., METALLINOU, A., BUSSO, C., LEE, S., AND NARAYANAN, S. S. *Speech in affective computing*. Oxford Univ. Press New York, NY, USA, New York, USA, 2014, p. 170–183.

[188] LEE, S., KIM, S., AND LEE, S. "What does your Agent look like?". In *Ext. Abstr. 2019 CHI Conf. Hum. Factors Comput. Syst.* (New York, NY, USA, may 2019), ACM, pp. 1–6.

[189] LIAO, Q. V., MAS-UD HUSSAIN, M., CHANDAR, P., DAVIS, M., KHAZAENI, Y., CRASSO, M. P., WANG, D., MULLER, M., SHAMI, N. S., AND GEYER, W. All work and no play? In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, Association for Computing Machinery, p. 1–13.

[190] LIAO, Q. V., AND MULLER, M. Human-AI Collaboration: Towards socially-guided machine learning. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[191] Liljedahl, M., and Fagerlönn, J. Methods for sound design. In *Proc. 5th Audio Most. Conf. A Conf. Interact. with Sound - AM '10* (New York, New York, USA, 2010), ACM Press, pp. 1–8.

[192] Lim, V., Jense, A., Janmaat, J., and Funk, M. Eco-feedback for non-consumption. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (New York, NY, USA, 2014), UbiComp '14 Adjunct, ACM, pp. 99–102.

[193] Limerick, H., Moore, J. W., and Coyle, D. Empirical evidence for a diminished sense of agency in speech interfaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2015), CHI '15, Association for Computing Machinery, p. 3967–3970.

[194] Liu, W. Natural user interface- next mainstream product user interface. In *2010 IEEE 11th International Conference on Computer-Aided Industrial Design Conceptual Design 1* (Yiwu, China, 2010), vol. 1, Institute of Electrical and Electronics Engineers, pp. 203–205.

[195] Longhurst, R. Semi-structured interviews and focus groups. *Key methods in geography 3*, 2 (2003), 143–156.

[196] López, G., Quesada, L., and Guerrero, L. A. Alexa vs. siri vs. cortana vs. google assistant: A comparison of speech-based natural user interfaces. In *Advances in Human Factors and Systems Interaction* (Cham, 2018), I. L. Nunes, Ed., Springer International Publishing, pp. 241–250.

[197] Luger, E., and Sellen, A. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 5 2016), CHI '16, Association for Computing Machinery, p. 5286–5297.

[198] Lugović, S., Dunder, I., and Horvat, M. Techniques and applications of emotion recognition in speech. In *2016 39th international convention on information and communication technology, electronics and microelectronics (mipro)* (Rijeka, Croatia, 2016), IEEE, p. 1278–1283.

[199] Luria, M., Reig, S., Tan, X. Z., Steinfeld, A., Forlizzi, J., and Zimmerman, J. Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents. *DIS 2019 - Proceedings of the 2019 ACM Designing Interactive Systems Conference* (2019), 633–644.

[200] Madaio, M. A., Yarzebinski, E., Kamath, V., Zinszer, B. D., Hannon-Cropp, J., Tanoh, F., Akpe, Y. H., Seri, A. B., Jasińska, K. K., and Ogan, A. Collective support and independent learning with a voice-based literacy technology in rural communities.

In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, apr 2020), ACM, pp. 1–14.

[201] MANSUR, D. L., BLATTNER, M. M., AND JOY, K. I. Sound graphs: A numerical data analysis method for the blind. *Journal of medical systems 9*, 3 (1985), 163–174.

[202] MARKOVÁ, I., AND LINELL, P. Coding elementary contributions to dialogue: Individual acts versus dialogical interactions. *Journal for the theory of social behaviour 26*, 4 (1996), 353–373.

[203] MAXQDA - DISTRIBUTION BY VERBI GMBH. MAXQDA The art of data analysis, 2021.

[204] MCCARTHY, J., AND WRIGHT, P. Technology as experience. *interactions 11*, 5 (2004), 42–43.

[205] MCGOOKIN, D., AND BREWSTER, S. *Earcons*. Logos Publishing House, Berlin, Germany., 2011. publisher: Logos Verlag.

[206] MCLEAN, G., AND OSEI-FRIMPONG, K. Hey Alexa . . . examine the variables influencing the use of artificial intelligent in-home voice assistants. *Comput. Human Behav. 99*, April (oct 2019), 28–37.

[207] MCTEAR, M. F., CALLEJAS, Z., AND GRIOL, D. *The conversational interface*, vol. 6. Springer, 2016.

[208] MEAD, G. In morris cw. *Mind, self, and society: From the standpoint of a social behaviorist* (1934).

[209] MELIKOGLU, M., LIN, C. S. K., AND WEBB, C. Analysing global food waste problem: pinpointing the facts and estimating the energy content. *Central European Journal of Engineering 3*, 2 (2013), 157–164.

[210] MENNICKEN, S., AND HUANG, E. M. Hacking the Natural Habitat: An In-the-Wild Study of Smart Homes, Their Development, and the People Who Live in Them. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7319 LNCS. 2012, pp. 143–160.

[211] MENNICKEN, S., VERMEULEN, J., AND HUANG, E. M. From today's augmented houses to tomorrow's smart homes: New directions for home automation research. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (New York, NY, USA, 2014), UbiComp 2014, ACM, pp. 105–115.

[212] MERLEAU-PONTY, M. Phenomenology of perception (da landes, trans.). *New York* (2012).

[213] METHFESSEL, B. Revis fachwissenschaftliche konzeption: Soziokulturelle grundlagen der ernährungsbildung. *Paderborner Schriften zur Ernährungs- und Verbraucherbildung, Band 7* (2005).

[214] MEUSER, M., AND NAGEL, U. Das experteninterview—konzeptionelle grundlagen und methodische anlage. *Methoden der vergleichenden Politik-und Sozialwissenschaft: neue Entwicklungen und Anwendungen* (2009), 465–479.

[215] MOLS, I., VAN DEN HOVEN, E., AND EGGEN, B. Informing Design for Reflection. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (New York, NY, USA, 10 2016), vol. 23-27-Octo, ACM, pp. 1–10.

[216] MORRIS, C. W. Foundations of the theory of signs. In *International encyclopedia of unified science*. Chicago University Press, 1938, pp. 1–59.

[217] MUJTABA, D. F., AND MAHAPATRA, N. R. Modeling the automation level of cyber-physical systems designed for food preparation. In *2019 9th International Symposium on Embedded Computing and System Design* (New York, NY, USA, 2019), ISED, IEEE, pp. 1–5.

[218] MUNTEANU, C., AND PENN, G. Speech and hands-free interaction: Myths, challenges, and opportunities. In *Conf. Hum. Factors Comput. Syst. - Proc.* (New York, New York, USA, 2018), vol. 2018-April, ACM Press, pp. 1–4.

[219] MURAD, C., AND MUNTEANU, C. "i don't know what you're talking about, halexa": The case for voice user interface guidelines. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (New York, NY, USA, 2019), CUI '19, Association for Computing Machinery.

[220] MURAD, C., MUNTEANU, C., COWAN, B. R., AND CLARK, L. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Comput. 18*, 2 (apr 2019), 33–45.

[221] MYERS, C., FURQAN, A., NEBOLSKY, J., CARO, K., AND ZHU, J. Patterns for how users overcome obstacles in Voice User Interfaces. In *Conf. Hum. Factors Comput. Syst. - Proc.* (New York, New York, USA, 2018), vol. 2018-April, ACM Press, pp. 1–7.

[222] MYERS, C. M. Adaptive suggestions to increase learnability for voice user interfaces. In *Proc. 24th Int. Conf. Intell. User Interfaces Companion - IUI '19* (New York, New York, USA, 2019), ACM Press, pp. 159–160.

[223] MYERS, C. M., FURQAN, A., AND ZHU, J. The impact of user characteristics and preferences on performance with an unfamiliar voice user interface. In *Conf. Hum. Factors Comput. Syst. - Proc.* (New York, New York, USA, 2019), ACM Press, pp. 1–9.

[224] MYNATT, E. D. Designing with Auditory Icons.

[225] NAKAJIMA, T., LEHDONVIRTA, V., TOKUNAGA, E., AND KIMURA, H. Reflecting human behavior to motivate desirable lifestyle. In *Proceedings of the 7th ACM conference on Designing interactive systems - DIS '08* (New York, New York, USA, 2008), ACM Press, pp. 405–414.

[226] NARUMI, T. Multi-sensorial virtual reality and augmented human food interaction. In *Proceedings of the 1st Workshop on Multi-sensorial Approaches to Human-Food Interaction* (2016), pp. 1–6.

[227] NASS, C., AND LEE, K. M. Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (2000), pp. 329–336.

[228] NESBITT, K. V., AND BARRASS, S. Evaluation of a multimodal sonification and visualisation of depth of market stock data. Georgia Institute of Technology.

[229] NEUMANN, A., ELBRECHTER, C., PFEIFFER-LESSMANN, N., KÕIVA, R., CARLMEYER, B., RÜTHER, S., SCHADE, M., ÜCKERMANN, A., WACHSMUTH, S., AND RITTER, H. J. "KogniChef": A cognitive cooking assistant. *KI-Künstliche Intelligenz 31*, 3 (2017), 273–281.

[230] NIEBORG, D. B., AND HELMOND, A. The political economy of facebook's platformization in the mobile ecosystem: Facebook messenger as a platform instance. *Media, Culture & Society 41*, 2 (2019), 196–218.

[231] NORMAN, D. A. Natural user interfaces are not natural. interactions 17, 3 (may 2010), 6-10. *URL: http://doi. acm. org/10.1145/1744161.1744163, doi 10* (2010), 1744161–1744163.

[232] NOTHDURFT, F., ULTES, S., AND MINKER, W. Finding appropriate interaction strategies for proactive dialogue systems – an open quest. In *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication* (Sweden, 2015), no. 110, Linköping Electronic Conference Proceedings, pp. 73–80.

[233] NOWACKI, C., GORDEEVA, A., AND LIZÉ, A.-H. Improving the usability of voice user interfaces: a new set of ergonomic criteria. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments: 9th International Conference, DUXU 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22* (2020), Springer, pp. 117–133.

[234] OBRIST, M., COMBER, R., SUBRAMANIAN, S., PIQUERAS-FISZMAN, B., VELASCO, C., AND SPENCE, C. Temporal, affective, and embodied characteristics of taste experiences: a framework for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), pp. 2853–2862.

[235] OOGJES, D., BRUNS, M., AND WAKKARY, R. *Lyssna: A Design Fiction to Reframe Food Waste*. Association for Computing Machinery, New York, NY, USA, 2016, p. 109–112.

[236] OOGJES, D., WAKKARY, R., AND ALONSO, M. B. Listening to the food:a design approach to food waste. *Research through Design Conference* (Apr 2019).

[237] OSWALD, D. Non-speech audio-semiotics: A review and revision of auditory icon and earcon theory.

[238] OVIATT, S. Multimodal interfaces for dynamic interactive maps. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 1996), CHI '96, Association for Computing Machinery, p. 95–102.

[239] OVIATT, S., COULSTON, R., AND LUNSFORD, R. When do we interact multimodally? cognitive load and multimodal communication patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (New York, NY, USA, 2004), ICMI '04, Association for Computing Machinery, p. 129–136.

[240] PAAY, J., NIELSEN, H., LARSEN, H., AND KJELDSKOV, J. Happy bits. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (New York, NY, USA, 9 2018), ACM, pp. 584–596.

[241] PARASURAMAN, A., AND COLBY, C. L. An updated and streamlined technology readiness index: Tri 2.0. *Journal of service research 18*, 1 (2015), 59–74.

[242] PARITOSH, K., KUSHWAHA, S. K., YADAV, M., PAREEK, N., CHAWADE, A., AND VIVEKANAND, V. Food waste to energy: an overview of sustainable approaches for food waste management and nutrient recycling. *BioMed Research International 2017* (2017).

[243] PARVIAINEN, E., AND SØNDERGAARD, M. L. J. Experiential Qualities of Whispering with Voice Assistants. *Conf. Hum. Factors Comput. Syst. - Proc.* (2020), 1–13.

[244] PEIRCE, C. S. *Peirce on signs: Writings on semiotic*. UNC Press Books, North Carolina, 1991.

[245] PÉNEAU, S., LINKE, A., ESCHER, F., AND NUESSLI, J. Freshness of fruits and vegetables: consumer language and perception. *British Food Journal 111*, 3 (mar 2009), 243–256.

[246] PENTINA, I., HANCOCK, T., AND XIE, T. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior 140* (2023), 107600.

[247] PETTERSSON, J. S., AND WIK, M. The longevity of general purpose wizard-of-oz tools. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction* (New York, NY, USA, 2015), OzCHI '15, ACM, p. 422–426.

[248] PETTY, R. E., AND BRINOL, P. *The Elaboration Likelihood Model*. SAGE, London, UK, Aug 2011, p. 224–245.

[249] PETTY, R. E., AND CACIOPPO, J. T. Source factors and the elaboration likelihood model of persuasion. *ACR North American Advances NA-11* (1984).

[250] PFEIFER, R., AND SCHEIER, C. *Understanding intelligence*. MIT press, 2001.

[251] PHAM, C., AND OLIVIER, P. Slice&dice: Recognizing food preparation activities using embedded accelerometers. In *European Conference on Ambient Intelligence* (Berlin, Heidelberg, 2009), vol. 5859 of *Lecture Notes in Computer Science*, Springer, pp. 34–43.

[252] PHILLIPS, E., OSOSKY, S., GROVE, J., AND JENTSCH, F. From Tools to Teammates: Toward the Development of Appropriate Mental Models for Intelligent Robots. *Proc. Hum. Factors Ergon. Soc. Annu. Meet. 55*, 1 (sep 2011), 1491–1495.

[253] PORCHERON, M., FISCHER, J. E., MCGREGOR, M., BROWN, B., LUGER, E., CANDELLO, H., AND O'HARA, K. Talking with conversational agents in collaborative action. In *companion of the 2017 ACM conference on computer supported cooperative work and social computing* (2017), pp. 431–436.

[254] PORCHERON, M., FISCHER, J. E., REEVES, S., AND SHARPLES, S. Voice Interfaces in Everyday Life. In *Proc. 2018 CHI Conf. Hum. Factors Comput. Syst. - CHI '18* (New York, New York, USA, 2018), vol. 2018-April, ACM Press, pp. 1–12.

[255] PORCHERON, M., FISCHER, J. E., AND SHARPLES, S. "Do animals have accents?": Talking with agents in multi-party conversation. In *Proc. ACM Conf. Comput. Support. Coop. Work. CSCW* (New York, New York, USA, 2017), ACM Press, pp. 207–219.

[256] PORPINO, G. Household food waste behavior: Avenues for future research. 41–51.

[257] POWERS, A., AND KIESLER, S. The advisor robot: Tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (New York, NY, USA, 2006), HRI '06, Association for Computing Machinery, p. 218–225.

[258] PROVOOST, S., LAU, H. M., RUWAARD, J., AND RIPER, H. Embodied conversational agents in clinical psychology: a scoping review. *Journal of medical Internet research 19*, 5 (2017), e151.

[259] RANASINGHE, N., CHEOK, A., NAKATSU, R., AND DO, E. Y.-L. Simulating the sensation of taste for immersive experiences. In *Proceedings of the 2013 ACM international workshop on Immersive media experiences* (2013), pp. 29–34.

[260] RECKI, L., ESAU-HELD, M., LAWO, D., AND STEVENS, G. Ai said, she said - how users perceive consumer scoring in practice. In *Proceedings of Mensch Und Computer 2023* (New York, NY, USA, 2023), MuC '23, Association for Computing Machinery, p. 149–160.

[261] RECKWITZ, A. Toward a theory of social practices. *European Journal of Social Theory 5*, 2 (may 2002), 243–263.

[262] Reckwitz, A. Grundelemente einer theorie sozialer praktiken. eine sozialtheoretische perspektive. *Zeitschrift für Soziologie 32*, 4 (2003), 282–301.

[263] Reeves, S. Some conversational challenges of talking with machines.

[264] Reeves, S. Conversation considered harmful? In *Proceedings of the 1st International Conference on Conversational User Interfaces* (New York, NY, USA, 2019), CUI '19, Association for Computing Machinery.

[265] Reeves, S., Porcheron, M., and Fischer, J. 'This is not what we wanted': Conversation designing with for voice interfaces. *Interactions 26*, 1 (dec 2019), 47–51.

[266] Research and Markets/Statista. Global smart kitchen appliances market size in 2020 and 2028 (in million US dollars), 2021.

[267] Riva, G. Is presence a technology issue? some insights from cognitive sciences. *Virtual reality 13*, 3 (2009), 159–169.

[268] Rogers, Y. Moving on from weiser's vision of calm computing: Engaging ubicomp experiences. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, P. Dourish and A. Friday, Eds., vol. 4206 LNCS of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 404–421.

[269] Rogers, Y., Connelly, K., Tedesco, L., Hazlewood, W., Kurtz, A., Hall, R. E., Hursey, J., and Toscos, T. Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4717 LNCS. 2007, pp. 336–353.

[270] Rohde, M., Brödner, P., Stevens, G., Betz, M., and Wulf, V. Grounded design - a praxeological is research perspective. *J. Infom. Tech. 32*, 2 (2017), 163–179.

[271] Rönnberg, N. Sonification for Conveying Data and Emotion. In *Audio Most. 2021* (New York, NY, USA, sep 2021), ACM, pp. 56–63.

[272] Røpke, I. Theories of practice—new inspiration for ecological economic studies on consumption. *Ecological economics 68*, 10 (2009), 2490–2497.

[273] Rossi, A., Moros, S., Dautenhahn, K., Koay, K. L., and Walters, M. L. Getting to know kaspar : Effects of people's awareness of a robot's capabilities on their trust in the robot. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (oct 2019), IEEE, pp. 1–6.

[274] Rozendaal, M. C., Van Beek, E., Haselager, P., Abbink, D., and Jonker, C. M. Shift and Blend: Understanding the hybrid character of computing artefacts on a tool-agent spectrum. In *HAI 2020 - Proc. 8th Int. Conf. Human-Agent Interact.* (New York, NY, USA, nov 2020), ACM, pp. 171–178.

[275] SAH, Y. J. Talking to a pedagogical agent in a smart tv: modality matching effect in human-tv interaction. *Behav. Inf. Technol. 40*, 3 (feb 2021), 240–250.

[276] SALSELAS, I., PENHA, R., AND BERNARDES, G. Sound design inducing attention in the context of audiovisual immersive environments. *Pers. Ubiquitous Comput. 25*, 4 (aug 2021), 737–748.

[277] SAMSUNG ELECTRONICS CO., LTD. Samsung introduces a one-stop shop that curates every step of your cooking journey, 2021.

[278] SATO, A., WATANABE, K., AND REKIMOTO, J. Mimicook: a cooking assistant system with situated guidance. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction* (New York, NY, USA, 2014), TEI '14, ACM, pp. 121–124.

[279] SCHAFFER, S., AND REITHINGER, N. Conversation is multimodal - Thus conversational user interfaces should be as well. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (New York, NY, USA, 2019), CUI '19, ACM.

[280] SCHANES, K., DOBERNIG, K., AND GÖZET, B. Food waste matters-a systematic review of household food waste practices and their policy implications. *Journal of Cleaner Production 182* (2018), 978–991.

[281] SCHATZKI, T. R. *Social practices: A Wittgensteinian approach to human activity and the social*. Cambridge University Press, 1996.

[282] SCHATZKI, T. R. *The site of the social: A philosophical account of the constitution of social life and change*. Penn State Press, 2002.

[283] SCHATZKI, T. R. *The timespace of human activity: On performance, society, and history as indeterminate teleological events*. Lexington Books, 2010.

[284] SCHMITT, A., ZIERAU, N., JANSON, A., AND LEIMEISTER, J. M. Voice as a contemporary frontier of interaction design. In *European Conference on Information Systems (ECIS).-Virtual* (2021).

[285] SCHRAMM, L. T., DUFAULT, D., AND YOUNG, J. E. Warning: This robot is not what it seems! exploring expectation discrepancy resulting from robot design. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (2020), pp. 439–441.

[286] SCHRÖDER, M. Emotional speech synthesis: A review. In *Seventh European Conference on Speech Communication and Technology* (Aalborg, Denmark, 2001), Citeseer.

[287] SCHULLER, B., BATLINER, A., STEIDL, S., AND SEPPI, D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication 53*, 9–10 (2011), 1062–1087.

[288] SCHULLER, B., RIGOLL, G., AND LANG, M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE international conference on acoustics, speech, and signal processing* (USA, 2004), vol. 1, IEEE, p. I–577.

[289] SCIUTO, A., SAINI, A., FORLIZZI, J., AND HONG, J. I. "hey alexa, what's up?": A mixed-methods studies of in-home conversational agent usage. In *Proceedings of the 2018 Designing Interactive Systems Conference* (New York, NY, USA, 2018), DIS '18, Association for Computing Machinery, p. 857–868.

[290] SEABORN, K., MIYAKE, N. P., PENNEFATHER, P., AND OTAKE-MATSUURA, M. Voice in human–agent interaction: a survey. *ACM Computing Surveys (CSUR) 54*, 4 (2021), 1–43.

[291] SEBERGER, J. S. Reconsidering the user in IoT: the subjectivity of things. *Pers. Ubiquitous Comput. 25*, 3 (2021), 525–533.

[292] SEIÇA, M., ROQUE, L., MARTINS, P., AND CARDOSO, F. A. Contrasts and similarities between two audio research communities in evaluating auditory artefacts. In *Proc. 15th Int. Conf. Audio Most.* (New York, NY, USA, sep 2020), ACM, pp. 183–190.

[293] SHELDON, K. M., ELLIOT, A. J., KIM, Y., AND KASSER, T. What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of personality and social psychology 80*, 2 (2001), 325.

[294] SHI, Y., YAN, X., MA, X., LOU, Y., AND CAO, N. Designing emotional expressions of conversational states for voice assistants: Modality and engagement. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), p. 1–6.

[295] SHNEIDERMAN, B. The limits of speech recognition. *Communications of the ACM 43*, 9 (2000), 63–65.

[296] SHOVE, E. *The Design of Everyday Life*. Cultures of Consumption Series. Berg Publishers, 2007.

[297] SHOVE, E., PANTZAR, M., AND WATSON, M. *The dynamics of social practice: Everyday life and how it changes*. Sage, 2012.

[298] SIMPSON, J. Are CUIs Just GUIs with Speech Bubbles? In *Proc. 2nd Conf. Conversational User Interfaces* (New York, NY, USA, jul 2020), ACM, pp. 1–3.

[299] SKJUVE, M., FØLSTAD, A., FOSTERVOLD, K. I., AND BRANDTZAEG, P. B. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies 149* (2021), 102601.

[300] SPLENDIDRESEARCH. Eine repräsentative Umfrage unter 1 . 058 Deutschen zu ihren Gewohnheiten nach dem Aufstehen.

[301] SPRADLEY, J. P. *Participant observation.* Waveland Press, Long Grove, IL, USA, 2016.

[302] SPROLL, S., PEISSNER, M., AND STURM, C. From product concept to user experience: exploring ux potentials at early product stages. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (2010), pp. 473–482.

[303] STEIN, S., AND MCKENNA, S. J. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous computing* (New York, NY, USA, 2013), UbiComp '13, ACM, pp. 729–738.

[304] STEINBERG, M. Line as super app: Platformization in east asia. *Social Media+ Society 6*, 2 (2020), 2056305120933285.

[305] STENMARCK, Â., JENSEN, C., QUESTED, T., MOATES, G., BUKSTI, M., CSEH, B., JUUL, S., PARRY, A., POLITANO, A., REDLINGSHOFER, B., ET AL. *Estimates of European food waste levels.* IVL Swedish Environmental Research Institute, 2016.

[306] STEVENS, G., BODEN, A., WINTERBERG, L., GÓMEZ, J. M., AND BALA, C. Digitaler konsum: Herausforderungen und chancen der verbraucherinformatik.

[307] STEVENS, G., ROHDE, M., KORN, M., WULF, V., PIPEK, V., RANDALL, D., AND SCHMIDT, K. Grounded design. a research paradigm in practice-based computing. *V. Wulf; V. Pipek; D. Randall; M. Rohde* (2018), 139–176.

[308] STRATEGY ANALYTICS. Strategy Analytics: Prime Day Smart Speaker Sales Boost Keeps Amazon Well Ahead of the Chasing Pack in Q3 2019, 2019.

[309] STREIJL, R. C., WINKLER, S., AND HANDS, D. S. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. 213–227.

[310] STRENGERS, Y., KENNEDY, J., ARCARI, P., NICHOLLS, L., AND GREGG, M. Protection, Productivity and Pleasure in the Smart Home. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 5 2019), ACM, pp. 1–13.

[311] SUTTON, D. *Sensible Objects : Colonialism, Museums and Material Culture.* Bloomsbury Academic, 2006.

[312] SUTTON, S. J. Gender ambiguous, not genderless: Designing gender in voice user interfaces (vuis) with sensitivity. In *Proceedings of the 2nd conference on conversational user interfaces* (2020), p. 1–8.

[313] SUTTON, S. J., FOULKES, P., KIRK, D., AND LAWSON, S. Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (2019), pp. 1–14.

[314] SVENNEVIG, J. *Språklig samhandling: innføring i kommunikasjonsteori og diskursanalyse*. Landslaget for norskundervisning, 2001.

[315] SYRDAL, D. S., OTERO, N., AND DAUTENHAHN, K. Video prototyping in human-robot interaction: Results from a qualitative study. In *Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction* (New York, NY, USA, 2008), ECCE '08, Association for Computing Machinery.

[316] TAYLOR, A. S., HARPER, R., SWAN, L., IZADI, S., SELLEN, A., AND PERRY, M. Homes that make us smart. *Personal and Ubiquitous Computing 11*, 5 (5 2007), 383–393.

[317] TERRY, G., HAYFIELD, N., CLARKE, V., AND BRAUN, V. Thematic analysis. In *The SAGE handbook of qualitative research in psychology*, C. Willig and W. S. Rogers, Eds. Sage, London, 2017, pp. 17–37.

[318] THIEME, A., COMBER, R., MIEBACH, J., WEEDEN, J., KRAEMER, N., LAWSON, S., AND OLIVIER, P. 'we've bin watching you' designing for reflection and social persuasion to promote sustainable lifestyles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), pp. 2337–2346.

[319] THYSSEN, G., AND GROSVENOR, I. Learning to make sense: interdisciplinary perspectives on sensory education and embodied enculturation, 2019.

[320] TORKKELI, K., MäKELä, J., AND NIVA, M. Elements of practice in the analysis of autoethnographical cooking videos. *J. Consum. Cult.* (2018).

[321] TSUI, A. B. Beyond the adjacency pair. *Language in Society 18*, 4 (1989), 545–564.

[322] TUOMELA, S., IIVARI, N., AND SVENTO, R. User values of smart home energy management system. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia* (New York, NY, USA, 11 2019), ACM, pp. 1–12.

[323] TURING, A. M. Computing machinery and intelligence. *Mind 59*, October (1950), 433–60.

[324] TURK, M. Multimodal interaction: A review. *Pattern recognition letters 36* (2014), 189–195.

[325] ULLRICH, D., BUTZ, A., AND DIEFENBACH, S. Who do you follow?: Social robots' impact on human judgment. *ACM/IEEE International Conference on Human-Robot Interaction* (2018), 265–266.

[326] VAN BOXSTAEL, S., DEVLIEGHERE, F., BERKVENS, D., VERMEULEN, A., AND UYTTENDAELE, M. Understanding and attitude regarding the shelf life labels and dates on pre-packed food products by belgian consumers. *Food Control 37* (2014), 85–92.

[327] VANNUCCI, E., ALTARRIBA, F., MARSHALL, J., AND WILDE, D. Handmaking food ideals: Crafting the design of future food-related technologies. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems* (2018), pp. 419–422.

[328] VELASCO, C., OBRIST, M., PETIT, O., AND SPENCE, C. Multisensory technology for flavor augmentation: a mini review. *Frontiers in psychology 9* (2018), 26.

[329] VERBEEK, P.-P. *What Things Do.* Penn State University Press, 4 2005.

[330] VIDGEN, H. A., AND GALLEGOS, D. Defining food literacy and its components. *Appetite 76* (may 2014), 50–59.

[331] VOLKEL, S. T., BUSCHEK, D., AND EIBAND, M. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. *Conf. Hum. Factors Comput. Syst. - Proc.* (2021).

[332] VÖLKEL, S. T., KEMPF, P., AND HUSSMANN, H. Personalised chats with voice assistants: The user perspective. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (New York, NY, USA, 2020), CUI '20, Association for Computing Machinery.

[333] VÖLKEL, S. T., SCHNEEGASS, C., EIBAND, M., AND BUSCHEK, D. What is "intelligent" in intelligent user interfaces? a meta-analysis of 25 years of iui. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2020), IUI'20, ACM, pp. 477–487.

[334] VTYURINA, A., AND FOURNEY, A. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, Association for Computing Machinery, p. 1–7.

[335] WAHLSTER, W., AND MAYBURY, M. An introduction to intelligent user interfaces. *RUIU, San Francisco: Morgan Kaufmann* (1998), 1–13.

[336] WAKEFIELD, A., AND AXON, S. 'i'm a bit of a waster': Identifying the enablers of, and barriers to, sustainable food waste practices. *Journal of Cleaner Production 275* (2020).

[337] WALKER, B. N., AND KRAMER, G. Ecological psychoacoustics and auditory displays: Hearing, grouping, and meaning making. In *Ecological psychoacoustics*. Brill, Leiden, Niederlande, 2004, pp. 149–174.

[338] WALLENBORN, G., AND WILHITE, H. Rethinking embodied knowledge and household consumption. *Energy Research & Social Science 1* (mar 2014), 56–64.

[339] WANG, C.-S., CHANG, Y.-F., AND CHENG, H. L. Intelligent Bathroom Lift. In *2019 IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)* (5 2019), IEEE, pp. 137–138.

[340] WANG, J., YANG, H., SHAO, R., ABDULLAH, S., AND SUNDAR, S. S. Alexa as coach: Leveraging smart speakers to build social agents that reduce public speaking anxiety. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, apr 2020), ACM, pp. 1–13.

[341] WANG, Q. J., MESZ, B., AND SPENCE, C. Assessing the impact of music on basic taste perception using time intensity analysis. In *Proc. 2nd ACM SIGCHI Int. Work. Multisensory Approaches to Human-Food Interact.* (New York, NY, USA, nov 2017), {MHFI} 2017, ACM, pp. 18–22.

[342] WARDE, A. Consumption and theories of practice. *Journal of Consumer Culture 5*, 2 (jul 2005), 131–153.

[343] WARDE, A. After taste: Culture, consumption and theories of practice. *Journal of Consumer Culture 14*, 3 (nov 2014), 279–303.

[344] WATZLAWICK, P., BAVELAS, J. B., AND JACKSON, D. D. *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. WW Norton & Company, 2011.

[345] WECHSUNG, I., AND NAUMANN, A. B. Evaluating a multimodal remote control: The interplay between user experience and usability. In *2009 International Workshop on Quality of Multimedia Experience* (New York, NY, USA, 2009), IEEE, pp. 19–22.

[346] WEISER, M., AND BROWN, J. S. The Coming Age of Calm Technology. In *Beyond Calculation*. Springer New York, New York, NY, 1997, pp. 75–85.

[347] WEISS, B., TROUVAIN, J., BARKAT-DEFRADAS, M., AND OHALA, J. J. *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*. Springer, Singapore, 2021.

[348] WEIZENBAUM, J. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM 9*, 1 (1966), 36–45.

[349] WENGER, E. Communities of practice: A brief introduction.

[350] WENGER-TRAYNER, É. *Communities of practice*. Cambridge University Press, 1999.

[351] WHITE, R. W. Skill discovery in virtual assistants. *Commun. ACM 61*, 11 (oct 2018), 106–113.

[352] WHITTINGTON, W. B. *Sound design and science fiction*. University of Southern California, New York, USA, 1999.

[353] WIENER, M., DEVOE, S., RUBINOW, S., AND GELLER, J. Nonverbal behavior and nonverbal communication. *Psychological review 79*, 3 (1972), 185.

[354] WILHITE, H. Towards a better accounting of the roles of body, things and habits in consumption. *Collegium 12* (2012), 87–99.

[355] WILLIAMS, H., WIKSTRÖM, F., OTTERBRING, T., LÖFGREN, M., AND GUSTAFSSON, A. Reasons for household food waste with special attention to packaging. *Journal of cleaner production 24* (2012), 141–148.

[356] WILSON, C., HARGREAVES, T., AND HAUXWELL-BALDWIN, R. Smart homes and their users: a systematic analysis and key challenges. *Personal and Ubiquitous Computing 19*, 2 (2 2015), 463–476.

[357] WINKLER, R., SÖLLNER, M., NEUWEILER, M. L., CONTI ROSSINI, F., AND LEIMEISTER, J. M. Alexa, can you help us solve this problem? In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, may 2019), ACM, pp. 1–6.

[358] WONG, B. W., AND BLANDFORD, A. Field research in hci: a case study. In *Proceedings of the 4th Annual Conference of the ACM Special Interest Group on Computer-Human Interaction* (2003), pp. 69–74.

[359] WOODWARD, K., AND KANJO, E. Things of the Internet (ToI). In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (New York, NY, USA, 10 2018), UbiComp '18, ACM, pp. 1228–1233.

[360] WOOLLEY, E., GARCIA-GARCIA, G., TSENG, R., AND RAHIMIFARD, S. Manufacturing resilience via inventory management for domestic food waste. *Procedia CIRP 40* (2016), 372–377.

[361] WULF, V., MÜLLER, C., PIPEK, V., RANDALL, D., ROHDE, M., AND STEVENS, G. *Practice-Based Computing: Empirically Grounded Conceptualizations Derived from Design Case Studies*. Springer, 2015, pp. 111–150.

[362] WULF, V., ROHDE, M., PIPEK, V., AND STEVENS, G. Engaging with practices. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11* (New York, New York, USA, 2011), ACM Press, p. 505.

[363] XU, Y., AND WARSCHAUER, M. What Are You Talking To?: Understanding Children's Perceptions of Conversational Agents. 1–13.

[364] YEWDALL, D. L. *Practical Art of Motion Picture Sound*. Taylor & Francis, Waltham, MA, USA, Aug 2012.

[365] YORKE-SMITH, N., SAADATI, S., MYERS, K. L., AND MORLEY, D. N. The design of a proactive personal agent for task management. *International Journal on Artificial Intelligence Tools 21*, 01 (2012).

[366] YU, Z., NICOLICH-HENKIN, L., BLACK, A. W., AND RUDNICKY, A. A Wizard-of-Oz Study on A Non-Task-Oriented Dialog Systems That Reacts to User Engagement. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Stroudsburg, PA, USA, 2016), Association for Computational Linguistics, pp. 55–63.