Digital Society (2023) 2:34
https://doi.org/10.1007/s44206-023-00062-2

ORIGINAL PAPER

Debiasing Strategies for Conversational AI: Improving Privacy and Security Decision-Making

Anna Leschanowsky[1] · Birgit Popp[1] · Nils Peters[2]

Received: 6 December 2022 / Accepted: 22 August 2023 / Published online: 9 September 2023

Abstract
With numerous conversational AI (CAI) systems being deployed in homes, cars, and public spaces, people are faced with an increasing number of privacy and security decisions. They need to decide which personal information to disclose and how their data can be processed by providers and developers. On the other hand, designers, developers, and integrators of conversational AI systems must consider users' privacy and security during development and make appropriate choices. However, users as well as other actors in the CAI ecosystem can suffer from cognitive biases and other mental flaws in their decision-making resulting in adverse privacy and security choices. Debiasing strategies can help to mitigate these biases and improve decision-making. In this position paper, we establish a novel framework for categorizing debiasing strategies, show how existing privacy debiasing strategies can be adapted to the context of CAI, and assign them to relevant stakeholders of the CAI ecosystem. We highlight the unique possibilities of CAI to foster debiasing, discuss limitations of the strategies, and identify research challenges.

Keywords Privacy · Security · Nudges · Debiasing · Decision-making · Conversational AI · CAI · Speech interfaces · Virtual voice assistant · VVA · Chatbots

✉ Anna Leschanowsky
anna.leschanowsky@iis.fraunhofer.de

Birgit Popp
birgit.popp@iis.fraunhofer.de

Nils Peters
nils.peters@fau.de

[1] Fraunhofer IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany

[2] International Audio Laboratories Erlangen, Am Wolfsmantel 33, 91058 Erlangen, Germany

1 Introduction

Interactions with conversational AI (CAI) systems become more and more widespread in everyday life. Virtual voice assistants offer hands-free communication in people's homes and cars and are increasingly deployed in public spaces such as in health institutions, accommodation places, or professional environments. Moreover, text-based systems are common to assist people in their online activities, e.g., in e-commerce or finance. To provide assistance, these systems ask people to disclose various personal information or request access to a wide range of personal data. In addition, inferences can be drawn from users' voice recordings or transcripts revealing sensitive information about themselves (Singh, 2019; Welch et al., 2019). To protect themselves from self-disclosure, users are faced with an increasing number of privacy and security decisions. They do not only need to decide whether to share information with a system but also in which way their information can be used and processed. However, given the complexity of the conversational AI ecosystem, it can be difficult for users to understand data-processing flows and possible implications to their privacy—a condition that is described as information asymmetry (Abdi et al., 2019; Acquisti et al., 2018). Moreover, information disclosure can be amplified as CAI systems aim to interact with users in a natural and human-like way and create an enjoyable and frictionless interaction (Seaborn et al., 2022). A positive mood can thereby influence people to underestimate privacy and security threats, increase the likelihood of disclosure, and serve as a mental shortcut (Alashoor et al., 2018; Dinev et al., 2015). In addition, peoples' privacy decision-making can suffer from varying systematic deviations in judgements, i.e., cognitive and behavioural biases (Acquisti et al., 2018).

While recognizing the importance of evaluation and mitigation of algorithmic biases for CAI systems (Beattie et al., 2022; Orphanou et al., 2022), in this paper, we focus on cognitive biases and their impact on human decision-making concerning privacy and security throughout the development and usage of CAI systems. Moreover, a large body of research has discussed privacy and security attacks, privacy risks for conversational AI systems and mitigation strategies from a technical point of view (Alepis & Patsakis, 2017; Bispham et al., 2022; Bispham et al., 2020; Pal et al., 2020). While technical safeguards are essential to ensuring privacy and security, it is equally important to support people in overcoming their biases to enable the implementation and usage of privacy-preserving techniques and non-regrettable privacy decision-making. Consequently, we take on a human-centric approach by focusing on the people involved in development, deployment, and usage of these systems and present strategies that support their privacy decision-making.

To understand and assist users' privacy and security decision-making, a growing body of research has applied behavioral economics (Acquisti et al., 2015; Ioannou et al., 2021). Design strategies that build on behavioral economic research aim to mitigate cognitive biases and improve users' privacy choices. One stream of behavioral economic research has explored nudging strategies to nudge users towards "better" decisions without restricting their options (Thaler & Sunstein, 2021). While nudging strategies for privacy decision-making have been successfully deployed and tested in the context of mobile applications, e-commerce, and social media (Acquisti et al., 2018; Almuhimedi et al., 2015; Ioannou et al., 2021; Wang et al., 2013), to the best of our knowledge, nudging strategies for CAI have been researched more generally but without specific focus on privacy choices (Zargham et al., 2022). However, as every design decision can influence users' choices for better or worse (Thaler & Sunstein, 2021), system providers and conversation designers have the potential to create a more private and secure experience by understanding users' cognitive and behavioral biases and by applying nudging techniques. In addition, cognitive flaws can prevent system providers and developers from designing, implementing, and deploying secure and private systems in the first place. Therefore, strategies based on behavioral economics can assist all actors in the ecosystem to make better judgements regarding privacy and security.

Nudging strategies are only one way to mitigate cognitive biases and support people's decision-making. In social science research, they are largely described as modifications to the environment (Soll et al., 2015). Yet, debiasing strategies can also focus on modifying a person's cognitive process. The medical field is especially rich in strategies that aim to mitigate biases and support decision-making through educational and cognitive strategies (Croskerry et al., 2013; Lambe et al., 2016). Therefore, we draw from previous categorizations of debiasing strategies to establish a novel categorization framework for debiasing techniques in the CAI context, adapt existing privacy debiasing techniques to conversational AI systems, and assign them to the relevant actors of the ecosystem. Therefore, in this position paper, we first provide an overview of actors of the CAI ecosystem in Section 4.1. While an extensive overview of biases and heuristics is out of the scope, we introduce the main sources of poor privacy and security decision-making in Section 4.2. We then establish a novel categorization framework, cluster debiasing strategies, and discuss their adaptation to CAI in Section 4.2.1. Based on our differentiation of actors in the conversational AI ecosystem, we introduce complementary strategies that can benefit various actors in the CAI ecosystem. Lastly, we discuss limitations and future research challenges in Section 4.2.3.

2 Conversational AI Ecosystem and Actors

Conversational AI can encompass a multitude of systems as it generally refers to technologies that allow natural interactions between machines and humans by leveraging AI-enabled speech and text processing (McTear, 2021). Thereby, conversational AI refers to text-based as well as voice-enabled systems. While voice-enabled systems require access to microphones and speakers, text-based systems rely on graphical user interfaces to allow input and output of messages. Moreover, multimodal applications are possible, e.g., voice assistants may be accompanied by a screen which allows displaying of complementary information. Both text as well as voice-based systems can be deployed on different physical instances such as on smart speakers, computers, or smart devices. In this position paper, we adopt a broader perspective by focusing on CAI systems as a whole considering their widespread application, multimodality, and growing relevance (e.g., OpenAI (2023)). All CAI systems, text-, and voice-based as well as multimodal systems share their conversational nature, i.e., users interact in natural language with them. By developing a framework for this broader scope that can be applied to all CAI systems, we hope to address challenges of debiasing human decision-making in CAI comprehensively. While it is possible that different types of CAI systems (e.g., text- vs. voice-based) may differ in the strategies that work best with them, a more general framework can be applied nonetheless, e.g., as guide to systematically test and compare strategies.

In their guidelines on virtual voice assistants (VVA), the European Data Protection Board (EDPB) (2021) identified relevant actors in the ecosystem. They differentiate between the VVA provider or designer, application developer, integrator, owner, and user. Figure 1 shows different actors and their individual tasks. We believe that their differentiation provides a suitable starting point to address relevant actors in the CAI ecosystem as the actor's task description is also applicable to the development of text-based CAI systems. Moreover, the number of defined roles is manageable and suitable for addressing the roles' hindrances in decision-making and recommending suitable debiasing strategies. Finally, the roles are not closely connected to legal definitions, e.g., data collector or processor under GDPR. Instead, due to the complexity of the CAI ecosystem, actors can take on different responsibilities or share data controlling

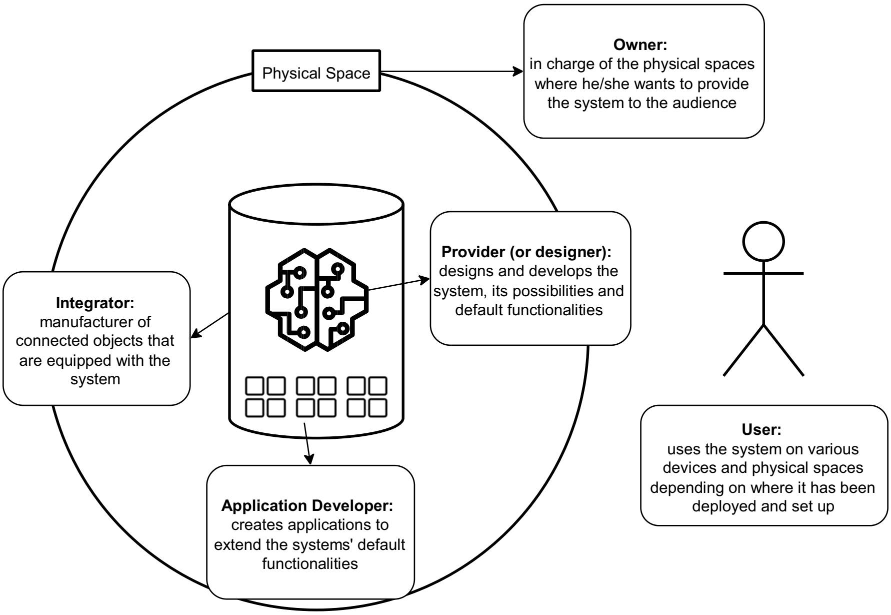
Physical Space
Owner: in charge of the physical spaces where he/she wants to provide the system to the audience
Provider (or designer): designs and develops the system, its possibilities and default functionalities
Integrator: manufacturer of connected objects that are equipped with the system
Application Developer: creates applications to extend the systems' default functionalities
User: uses the system on various devices and physical spaces depending on where it has been deployed and set up

Fig. 1 Differentiation of actors in the CAI ecosystem as described by (European Data Protection Board (EDPB), 2021)

(European Data Protection Board (EDPB), 2021; van Mil & Quintais, 2022). As a legal assessment is out of scope of this article, we will rely only on the actors' tasks as shown in Fig. 1 to map suitable debiasing strategies. In general, debiasing strategies can be applied independently of rights and obligations under legal regulations contributing to the overall principle of privacy by design and human's right to privacy (Cavoukian, 2009; United Nations, 1948). Moreover, individuals and organizations might not only inhabit one role but multiple roles. This is emphasized by Fig. 1 as it shows the close relationship between the provider, integrator, and application developer to successfully create a CAI system. For instance, businesses can act as providers and integrators by developing the system's main functionalities and manufacturing necessary components, e.g., a smart speaker. In addition, businesses can provide platforms that allow application developers to enhance the default functionality of the system, e.g., Google Actions or Alexa Skills. Even if inhabiting more than one role, people and organisations can profit from an overview that shows which debiasing strategies can be applied to these individual roles.

Importantly, the owner role can differ from the user role as CAIs are established in accommodation places or professional environments such as office spaces or schools. For example, when a system is deployed at a workplace, the company may be the owner of the system while its users are employees. Similarly, when the system is used in a school, the school may be the owner while teachers and students are its users. However, teachers may also inherit the role of an owner as they might be in charge of a single classroom where the conversational AI system is provided to students.

So far, nudges for privacy and security decision-making have largely focused on the relationship between online services and their end-users (Ioannou et al., 2021; Kitkowska et al., 2020). However, given the complexity of conversational AI ecosystems and the different actors involved in creating these experiences, we believe that nudges and debiasing strategies in general can provide helpful tools for fostering privacy and security solutions throughout the ecosystem.

3 Sources of Poor Privacy and Security Decision-Making

Different actors might suffer from varying hindrances and biases to make desirable privacy and security decisions. While insufficient access to information can be one of the major factors for users to engage in adverse privacy decisions, it might be less of a concern for application developers. They are involved in data usage, processing, and storage and therefore have access to more information than users. While a complete analysis of the varying hindrances of optimal decision-making for the individual stakeholders is out of scope, we identify some major hurdles in the following sections. This will allow us to recommend and suggest potentially helpful interventions for the different groups.

3.1 Information Asymmetry

In the field of privacy and security, it is common for users to be subject to incomplete or asymmetric information as data collectors usually have more or better information about the collection, processing, and storage of data (Acquisti et al., 2020). Therefore, users may be unable to make informed decisions about their privacy, e.g., whether to disclose information to a CAI system. Informed decision-making can be particularly difficult as users might be unaware of the existence of different actors and their data collection and processing, e.g., third parties (Abdi et al., 2019). Moreover, due to the power imbalance, users can become subject to persuasive conversations with possible consequences to their privacy and security (Murtarelli et al., 2021).

Professionals involved in building or establishing CAI systems form a large group (European Data Protection Board (EDPB), 2021). They can include owners, integrators, application developers, and providers—and all of them may be more knowledgeable than users about CAI systems' capabilities. However, technical expertise may vary within a role. For example, people with different backgrounds may be owners of CAI systems. While a teacher may have little technical background and may experience similar power imbalances as users, a manager with engineering background who decides to roll out CAI in the workplace may have a better understanding of the CAI ecosystem and its accompanied privacy and security risks. While app developers can be considered data collectors or processors when accessing or requesting certain personal attributes from the users, they might have limited knowledge about data processing done by conversational AI system providers. For instance, Amazon Alexa voice recordings are not shared with skill developers (Amazon Inc., 2019). In contrast, providers can have incomplete information about how data is handled by the application developers and whether personal attributes are directly requested (Lentzsch et al., 2021). This suggests that information asymmetry matters for all actors in the conversational AI ecosystem even though its impact might differ.

3.2 Heuristics and Biases

Privacy and security decision-making is often subject to uncertainty as long-term risks can be unknown or difficult to grasp and decisions are constrained by time and available information (Acquisti et al., 2015; Leschanowsky et al., 2021). Moreover, assessing the probability of possible malicious events and privacy breaches can be effortful and exhausting. The fact that human decision-making is subject to limited resources was first discussed by Simon (1990) under the concept of "bounded rationality." He pointed out that decision-makers may rely on heuristics or mental shortcuts to simplify the process. Later, Kahneman (2011) extended this idea by developing the dual-process model of cognition—the distinction between system 1 and system 2. System 1 refers to intuitive, fast, and effortless thinking which can, however, result in biased and suboptimal choices. In contrast, system 2 describes a slower, more conscious, and controlled thinking process which is likely to be beneficial for making complex privacy choices.

For an extensive overview of heuristics and biases that can impact privacy decisions, we refer readers to Acquisti et al., (2018, 2020). While research on cognitive biases that impact privacy decision-making has merely focused on online environments, a majority of biases is applicable to the context of conversational AI systems and their actors. For example, anthropomorphism, i.e., the perceived level of human-like characteristics, and other salient cues can influence users' intentions, disclosure, and privacy concerns (Cai et al., 2022; Ha et al., 2021; Ischen et al., 2020). Thereby, self-disclosure of CAI systems could effectively encourage users to share personal information (Rao et al., 2022). Moreover, mass media portrays and brand loyalty can influence privacy perceptions and usage of CAI systems (Maroufkhani et al., 2022; Sin et al., 2021). As a detailed analysis of specific biases of CAI systems is out of scope, we will rely on biases that are known to influence privacy and security decision-making for the remainder of this paper. Nevertheless, we emphasize the need to further investigate cognitive and behavioral biases that are specific or amplified in conversational AI systems, e.g., machine heuristic (Sundar & Kim, 2019).

3.3 Decision Readiness

Lastly, we want to focus on decision readiness that can negatively influence privacy decision-making in CAI. Decision readiness refers to the fact that system 2—the slow and more controlled thinking—is ready to monitor and if necessary intervene in intuitive thinking (Soll et al., 2015). However, this capability can be impeded by factors like fatigue, distraction, visceral influences, and individual differences. As CAI systems may be particularly useful when hands-free interaction is necessary or multiple tasks are carried out, e.g., in a car, factors like distraction constitute a crucial source of biased decision-making. As previously discussed, visceral influences might include visual or auditory cues of conversational AI systems which can lead to increased self-disclosure or suboptimal privacy decisions (Ischen et al., 2020). Lastly, individual differences such as differences in training, cognitive ability, or self-reflection can impact privacy decision-making. Therefore, debiasing strategies might show different levels of effectiveness depending on an individual's characteristics and role in the CAI ecosystem.

4 Categorizations of Debiasing Strategies and Their Limitations

An increased understanding of cognitive biases has spurred the development of bias mitigation strategies across various domains, including healthcare, finance and privacy, and security. Additionally, efforts were taken to categorize individual debiasing strategies into high-level classes, although categorizations differ within and between disciplines. For instance, in the medical field, debiasing strategies have been grouped into cognitive, technological, motivational, and affective strategies (Broussard & Wulfert, 2019; Larrick, 2004; Ludolph & Schulz, 2018). In contrast, other studies have distinguished between educational and workplace strategies as well as forcing functions (Croskerry et al., 2013; Lambe et al., 2016; Neal & Brodsky, 2016). The former categorization is based on assumptions of different strategies on how to approximate ideal decision-making outcomes (Larrick, 2004), while the latter distinguishes based on the temporal appearance of a debiasing effect with educational strategies influencing future decision-making and workplace strategies helping to overcome bias at the time of decision-making without necessarily changing the individual (Croskerry et al., 2013). Interestingly, medical research has predominantly focused on strategies that modify the person, e.g., through cognitive training (Lambe et al., 2016; Ludolph & Schulz, 2018), whereas research on debiasing in the privacy and security domain has primarily concentrated on the implementation and evaluation of nudges, e.g., nudging with information and presentation, defaults, or incentives (Acquisti et al., 2018; Ioannou et al., 2021; Kitkowska et al., 2020). In general, the term "nudge" stems from behavioral economics and describes "any aspect of the choice architecture that alters people's behavior predictably without forbidding any options or significantly changing their economic incentives" (Thaler & Sunstein, 2021). It can be used by "choice architects" to influence decision-making by modifying the environment (Thaler & Sunstein, 2021). The term "nudges" can also be utilized as an acronym to cluster different interventions, i.e., iNcentives, Understand mappings, Defaults, Give feedback, Expect errors, Saliency (Acquisti et al., 2018). However, due to the nature of CAI systems, their seamless way of interacting with users via natural language, and their complex ecosystems, there is a need to move beyond the application of nudges for privacy and security decision-making, to draw from other disciplines, and to establish a more comprehensive categorization framework. While a universally accepted taxonomy for debiasing strategies is yet to be established, Soll et al. (2015) proposed a more broadly applicable categorization distinguishing strategies that modify the person to those that modify the environment. Their framework has been applied in diverse research fields such as geoscience education and management decision-making (Muntwiler, 2021; Wilson et al., 2019). In particular, Muntwiler (2021) illustrate how debiasing strategies can be theoretically grouped into a two-level categorization framework following Larrick (2004) and Soll et al. (2015) with modifications to the person and the environment building higher-level categories.

4.1 Towards a Categorization Framework for Debiasing Strategies for Conversational AI Systems

Given previous research on classification logics, we identify two high-level categorization frameworks to cluster debiasing strategies for CAI systems, i.e., categorization depending on the temporal appearance of the debiasing effect (Croskerry et al., 2013) and categorization depending on the type of modification (Soll et al., 2015). Due to their simplicity and clarity, they can help to navigate the landscape of debiasing strategies for CAI and provide a starting point for more detailed and nuanced frameworks. While each framework individually offers valuable insights, we recognize their complementary perspectives and their potential to capture the underlying principles and patterns of debiasing strategies for CAI. Being derived from the medical field with a strong focus on cognitive interventions, the distinction between educational and workplace strategies might overlook the impact of environmental modifications on future or real-time decision-making. Likewise, distinguishing only between modifications to the individual and the environment does not adequately address the benefits of CAI systems in seamlessly interacting with users and their potential to debias human privacy and security decisions. For instance, CAI systems can proactively function as guides, mentors, or teachers on privacy, security, and protective mechanisms and thereby modifying both the environment and the individual. Combining these two frameworks enables a comprehensive approach, integrating complementary perspectives while maintaining clarity and simplicity in the classification of debiasing strategies for CAI systems.

We propose a two-dimensional categorization framework for debiasing strategies in the context of privacy and security for CAI. Thereby, Fig. 2 facilitates a better understanding of the relationships between the frameworks, their compatibility, and usefulness for consolidating approaches from various disciplines. By establishing a

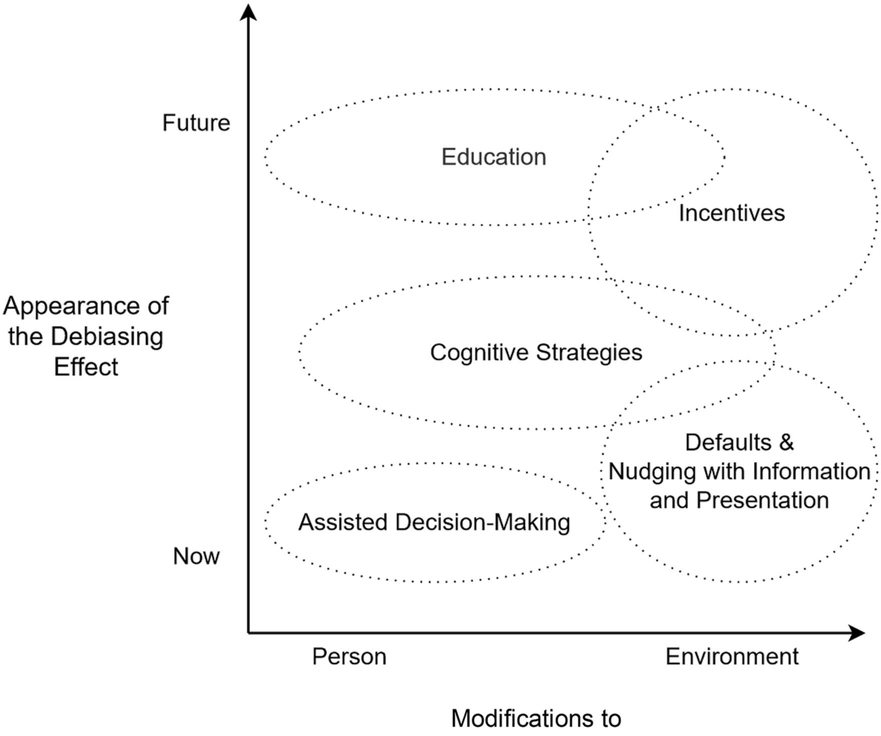

Fig. 2 Schematic representation of the categorization framework for debiasing strategies for conversational AI. We rely on previous categorization logics and research to define the placement of classes (Acquisti et al., 2018; Croskerry et al., 2013; Soll et al., 2015). By using dotted lines, we emphasize that the expansion of the classes is based on our assessment and might vary depending on the debiasing strategies considered for a certain context

graph visualization, we do not consider debiasing classes to be mutually exclusive. Thus, we acknowledge that certain debiasing strategies, i.e., cognitive strategies, may lie in the area between modifications to the person and the environment as well as affecting peoples' decision-making now and in the future.

For the remainder of this paper, we will particularly focus on five high-level classes, i.e., education, incentives, cognitive strategies, assisted decision-making, defaults, and nudging with information and presentation. While previous work has introduced the class of automated decision-making or technological strategies (Larrick, 2004; Soll et al., 2015), we will refer to this class as assisted decision-making to emphasize the role of humans in the decision-making process. The classes were chosen as they are reoccuringly discussed in previous research in various fields (Acquisti et al., 2018; Croskerry et al., 2013; Soll et al., 2015) and their benefits of complementing each other. For example, while education focuses on modifying the person and their future decision-making, information and presentation provision refers to a change in the environment that can influence people at the time of decision-making. On the other hand, incentives can be considered modifications to the environment with mid-term to long-term influence on peoples' decision-making, while assisted decision-making such as privacy assistants modify the person and their decision in a specific moment. Our framework emphasizes that the chosen strategies need to be seen as complementary to each other, as all these interventions come with strengths and weaknesses.

4.2 Debiasing Strategies

We will now extend on the previously introduced classes by presenting corresponding debiasing strategies adaptable to CAI and their limitations. For each class, we provide an overview of discussed strategies through the use of tables (see Tables 1, 2, 3, 4, 5, and 6). Whenever applicable, we point to references that discuss the proposed strategies in the context of CAI or related technologies, e.g., app development, and show for which actors they have been applied. We deliberately leave cells blank where the proposed strategy has—to the best of our knowledge—not yet been studied in-depth for a specific actor in CAI or related ecosystems, emphasizing the need for future research in these areas. Thereby, we aim to draw attention to promising opportunities for further investigation.

Lastly, we want to emphasize that we present a first attempt of mapping existing debiasing strategies to CAI and that we do not provide a comprehensive overview of available debiasing strategies. Instead, we bring together strategies that have been applied for privacy and security decision-making in online environments and effective strategies from other disciplines. Thereby, we focus on expanding the range of possible debiasing strategies for CAI while keeping them applicable to the various actors of the ecosystem.

Table 1 Overview of debiasing strategies for the class of "education" for ensuring privacy and security in conversational AI systems. Whenever applicable, we point to references that discuss these strategies in the context of CAI or related technologies. We show them in the actors' columns that are directly addressed by those references or closely connected (see Section 2 for a detailed description of the actors). For greater clarity, we have restricted the choice of references to one and will discuss further examples in the associated sections. We deliberately leave cells blank where the proposed strategy has—to the best of our knowledge—not yet been studied in-depth for a specific actor in CAI or related ecosystems, emphasizing the need for future research in these areas

Debiasing Strategy	Provider (or Designer)	App Developer	Integrator	Owner	User
Guided Reflection					✓[a]
Technical Guidelines Catalog		✓[b]			
Privacy Patterns	✓[c]	✓[c]			

[a] Kocielnik et al. (2018)

[b] Hatamian (2020)

[c] UC Berkeley School of Information (2019)

4.2.1 Education

Education for Debiasing Decision-Making One way to improve decision-making is to educate individuals and increase their competencies in privacy and data protection over time. Studies have shown that a higher level of online privacy literacy

Table 2 Overview of debiasing strategies for the class of "cognitive strategies" for ensuring privacy and security in conversational AI systems. Whenever applicable, we point to references that discuss these strategies in the context of CAI or related technologies. We show them in the actors' columns that are directly addressed by those references or closely connected (see Section 2 for a detailed description of the actors). For greater clarity, we have restricted the choice of references to one and will discuss further examples in the associated sections. We deliberately leave cells blank where the proposed strategy has—to the best of our knowledge—not yet been studied in-depth for a specific actor in CAI or related ecosystems, emphasizing the need for future research in these areas

Debiasing Strategy	Provider (or Designer)	App Developer	Integrator	Owner	User
Generating Alternatives					✓[a]
Cognitive Forcing					✓[b]
Active Choice					✓[c]
Prospective Hindsight					
Strategies Increasing the Accuracy of Judgements					
Planned Interruptions					✓[d]
Planning Prompts					✓[e]

[a] Bach et al. (2023)

[b] Bucinca et al. (2021)

[c] Utz et al. (2019)

[d] Wang et al. (2013)

[e] Cuadra et al. (2021)

Table 3 Overview of debiasing strategies for the class of "assisted decision-making" for ensuring privacy and security in conversational AI systems. Whenever applicable, we point to references that discuss these strategies in the context of CAI or related technologies. We show them in the actors; columns that are directly addressed by those references or closely connected (see Section 2 for a detailed description of the actors). For greater clarity, we have restricted the choice of references to one and will discuss further examples in the associated sections. We deliberately leave cells blank where the proposed strategy has—to the best of our knowledge—not yet been studied in-depth for a specific actor in CAI or related ecosystems, emphasizing the need for future research in these areas

Debiasing Strategy	Provider (or Designer)	App Developer	Integrator	Owner	User
Privacy Assistant					✓[a]
Checklists	✓[b]				

[a] Colnago et al. (2020)

[b] Open Voice Network (2023)

can significantly increase the usage of protective strategies (Masur et al., 2017; Park, 2013). Thereby, most educational strategies aim at modifying the person and impacting their future decision-making capabilities. Privacy literacy can be divided into factual and procedural knowledge (Masur, 2019). While factual knowledge refers to expertise on certain technical or legal aspects regarding data protection and processing, procedural knowledge is concerned with the ability to use protective privacy strategies (Masur, 2019). While both aspects are essential for all actors in the conversational AI ecosystem, there might be certain priorities set for different stakeholders. For example, educational strategies for users might emphasize procedural knowledge while strategies that focus on providers and app developers need to stress knowledge of technical and legal aspects. Moreover, users might benefit the most from procedural knowledge that allows them to apply strategies for individual protection, while providers and app developers have to be knowledgeable about strategies for users' privacy protection.

Table 4 Overview of debiasing strategies for the class of "incentives" for ensuring privacy and security in conversational AI systems. Whenever applicable, we point to references that discuss these strategies in the context of CAI or related technologies. We show them in the actors' columns that are directly addressed by those references or closely connected (see Section 2 for a detailed description of the actors). For greater clarity, we have restricted the choice of references to one and will discuss further examples in the associated sections. We deliberately leave cells blank where the proposed strategy has—to the best of our knowledge—not yet been studied in-depth for a specific actor in CAI or related ecosystems, emphasizing the need for future research in these areas

Debiasing Strategy	Provider (or Designer)	App Developer	Integrator	Owner	User
Badges and App Reviews					✓[a]
Organizational Measures					
Regulations	✓[b]				

[a] Acquisti et al. (2018)

[b] Kosseff (2016)

Table 5 Overview of debiasing strategies for the class of "defaults" for ensuring privacy and security in conversational AI systems. Whenever applicable, we point to references that discuss these strategies in the context of CAI or related technologies. We show them in the actors' columns that are directly addressed by those references or closely connected (see Section 2 for a detailed description of the actors). For greater clarity, we have restricted the choice of references to one and will discuss further examples in the associated sections. We deliberately leave cells blank where the proposed strategy has—to the best of our knowledge—not yet been studied in-depth for a specific actor in CAI or related ecosystems, emphasizing the need for future research in these areas

Debiasing Strategy	Provider (or Designer)	App Developer	Integrator	Owner	User
Default Settings		✓[a]			✓[b]

[a] Amazon Inc. (2022b)

[b] Lau et al. (2018)

As privacy and security become an essential part of technological systems, the need for education and training is growing as well as the diversity of resources for privacy education. An interview study by Subramaniam et al. (2019) revealed that people rely on different sources to educate themselves on the topic of privacy. These include school and educational lessons and job training, experiences, and knowledge from family members and friends and so-called privacy experts, i.e., bank employees or librarians as well as popular organizations. Moreover, experiences with privacy issues and system-programmed privacy measures, e.g., regular prompts to change passwords, played a crucial role in engaging in protective behavior (Subramaniam et al., 2019). On the other hand, studies have shown that certain groups might suffer from lower levels of privacy literacy and may be less likely to apply protective strategies. For example, sociodemographic factors, e.g.,

Table 6 Overview of debiasing strategies for the class of "nudging with information and presentation" for ensuring privacy and security in conversational AI systems. Whenever applicable, we point to references that discuss these strategies in the context of CAI or related technologies. We show them in the actors' columns that are directly addressed by those references or closely connected (see Section 2 for a detailed description of the actors). For greater clarity, we have restricted the choice of references to one and will discuss further examples in the associated sections. We deliberately leave cells blank where the proposed strategy has—to the best of our knowledge—not yet been studied in-depth for a specific actor in CAI or related ecosystems, emphasizing the need for future research in these areas

Debiasing Strategy	Provider (or Designer)	App Developer	Integrator	Owner	User
Privacy Labels					✓[a]
Choice Engines					✓[b]
Conversational Privacy					✓[c]
Warnings and Reminders		✓[d]			✓[e]

[a] Emami-Naeini et al. (2020)

[b] Mozilla (2022)

[c] Harkous et al. (2016)

[d] Tahaei et al. (2021)

[e] Yeasmin et al. (2020)

income, education, age, and gender, can influence the level of privacy literacy (Park, 2013). Children form a particularly vulnerable group as they are not yet literate in privacy risks but are exposed to CAI systems through children's toys and smart home technologies (Mann et al., 2022). Children's usage of IoT devices and smart speakers and their protection largely depends on their guardians who come with significant differences in their level of privacy literacy and concerns, e.g., due to gender, racial, and socioeconomic differences (Garg & Sengupta, 2019; Mann et al., 2022). Nevertheless, when it comes to voice recordings, people are in general largely unaware of inferences that can be drawn from their voices (Kröger et al., 2022). This shows a need for easily useable and adaptive educational programs to raise awareness of privacy risks of CAI systems for various user groups.

Educational Debiasing Strategies in CAI While educational strategies have traditionally focused on modifying the person through training courses, tutoring, or similar interventions (Croskerry et al., 2013), due to their human-like capabilities, CAI systems can proactively trigger educational interventions taking into account individual differences and context. Proactive educational approaches can be purposefully designed and therefore constitute not only to modifications to the person but also to the environment. Thereby, CAI systems can act as guides and mentors to people to raise awareness and promote privacy literacy (Leschanowsky et al., 2022). In previous work, we showed how the concept of *Guided Reflection*—a strategy that has been successfully applied in the medical context to increase diagnostic accuracy through mentoring and self-reflection—can be leveraged by CAI systems (Leschanowsky et al., 2022). In fact, conversational agents have been successfully used to support workers in their self-reflection and self-learning, e.g., by prompting workers to journal activities (Kocielnik et al., 2018). As children form a particularly vulnerable group, providing suitable education and mentoring on privacy aspects through CAI can significantly increase the young generation's privacy literacy. Thereby, design recommendations on learning applications for children in CAI (Garg & Sengupta, 2020) can inform the development of educational privacy tools (Table 1).

While most of the discussion above focused on user's privacy literacy and its limitations, boosting provider's, app developer's, and owner's privacy literacy (see Fig. 1) is key to creating private and secure CAI systems. However, recent studies found that privacy is not considered thoughtfully throughout development. Edu et al. (2022) investigated voice applications on the Alexa Marketplace and found that 36% of skills follow bad privacy practices such as broken traceability, i.e., the privacy policy does not cover data practices. Moreover, Liao et al. (2020) could show that current privacy policies of voice applications are often non-existent, incorrect, or inaccessible.

To counter bad practices among mobile app developers, Hatamian (2020) designed a *Technical Guidelines Catalog* by mapping legal principles of the General Data Protection Regulation (GDPR) to technical privacy and security solutions. Such a catalog could also assist CAI application developers in integrating privacy during application design and development. Moreover, Privacy-by-Design offers a more extensive approach to integrate privacy into a system throughout the entire development lifecycle (Cavoukian, 2009). Therefore, *privacy patterns* provide concrete tools for common privacy problems to ensure privacy-friendly systems (UC Berkeley School of Information, 2019). Technical privacy and security solutions derived from legal principles may overlap with privacy patterns, but privacy patterns can be more diverse and do not necessarily map to certain legal requirements. While privacy patterns can benefit providers, developers, and integrators of conversational AI systems, pattern catalogs often lack consistency and are scattered among various platforms (Colesky et al., 2018). Moreover, while many of these patterns are applicable in the context of conversational AI systems, patterns matching the unique context of CAI are missing. For instance, human–computer interaction (HCI) privacy patterns rely mostly on visual cues and graphical interfaces (UC Berkeley School of Information, 2019). Yet, work by Murad et al. (2021) shows how grounding design guidelines for CAI on existing GUI heuristics can benefit adoption and how to transition between them. Future research could explore similar approaches for privacy patterns.

Limitations While education is a necessary and undeniable strategy to support people in making better privacy decisions, it is insufficient to fully mitigate biases. One reason is that it remains unclear how much of a difference privacy literacy makes and whether people can apply their skills in concrete situations. Fernandes et al. (2014) found that the efficacy of financial literacy training is modest while training effects were larger for students who were trained over longer periods. However, as the acquired financial literacy declined over time, the authors conclude that the most effective form of training is education that is provided at times when needed. Similarly, users who had additional training on privacy literacy might not experience a long-lasting effect. Therefore, complementary debiasing strategies that modify the environment by informing people at the time of decision-making are crucial and will be discussed in detail in Section 4.2.6.

While most of the proposed educational methods have focused on users or software engineers, little is known about educational methods for owners of conversational AI systems. While users interact with the system on a one-to-one basis, an owner is in charge of the physical space and does not necessarily need to interact with the system directly. Nevertheless, the owner needs to be aware of possible interconnectivity of the system and its data accessibility. For instance, when a system is deployed in an accommodation space, safeguards need to be taken such that guests cannot access data that has been provided by former guests. Moreover, in educational settings, teachers might take on an owner role if they are in charge of a classroom where the system is deployed. As children are considered a vulnerable group and their recordings might be sensitive, teachers need to undergo specific training to ensure that appropriate security and privacy measures are taken and the systems are used appropriately (Terzopoulos & Satratzemi, 2020). As these systems become widespread in accommodation, professional, and educational places, educating owners on privacy and data protection is crucial to ensure users' privacy and their acceptance of the technology.

Takeaway

Educational strategies can support all actors in the CAI ecosystem. CAI systems can leverage guided reflection to make users reflect and learn about their privacy and security decisions. Providers, application developers, and integrators can benefit from technical guideline catalogues and privacy patterns tailored for conversational AI systems. Lastly, there is a need to explore educational methods for owners of CAI systems.

4.2.2 Cognitive Strategies

Cognitive Strategies for Debiasing Decision-Making While education is concerned with boosting privacy literacy over time, cognitive strategies aim to impact people's cognitive abilities at the time of decision-making. Yet, cognitive strategies have also been used in the medical field to enhance decision-making over time and have the potential to create long-lasting effects on improved decision-making (Croskerry, 2003). While a variety of cognitive strategies has been tested in various fields, e.g., medical field, AI overreliance, privacy, and security (Bucinca et al., 2021; Croskerry, 2003; Wang et al., 2013), they are yet to be investigated in the field of CAI. Therefore, in this position paper, we focus on cognitive strategies that have been frequently discussed in previous research on debiasing (Croskerry et al., 2013; Larrick, 2004; Soll et al., 2015). Cognitive strategies can ask people to identify situations in which decision-making errors are likely to occur and deliberately apply strategies to avoid decision errors (Croskerry, 2003). Others aim at directly inducing reflection and asking individuals to engage their system 2 thinking capabilities through interruptions or specific ways of presenting choices. While these strategies can be triggered by a CAI system and therefore constitute to modifications to the environment, they are likely to modify the person by altering their thinking process (see Fig. 2 for the classification of cognitive strategies into the framework).

Cognitive Debiasing Strategies in CAI *Generating alternatives* and evaluating them based on established decision criteria is crucial for making rational decisions. However, due to cognitive biases, people are unlikely to engage in rational thinking and might have difficulties in generating alternatives (Soll et al., 2015). "Consider the Opposite" can be seen as a related strategy and has been proven helpful in clinical AI support (Bach et al., 2023). Due to their unique possibility of interacting with users naturally, CAI could support them in generating alternatives based on their own decision objectives. As privacy objectives can be highly subjective and dependent on people's attitudes and values, system designers and developers might have difficulties in sensibly curating alternatives for the users. Therefore, CAI systems can foster a rational decision-making process by having users list their decision criteria, e.g., privacy concerns or interest in using the service, and weigh them according to their importance. Moreover, CAI could assist users in generating alternatives and finding their optimal choice. While research has shown that generating alternatives is most problematic and difficult for humans (Nutt, 2004), CAI systems might be capable of generating alternatives in a fast, effective, and comprehensive way. This can make it easy for users to choose among a few alternatives that have been found to fit their decision criteria best. Engaging in a rational decision process can be especially useful in situations where decisions are complex and only need to be made once, e.g., deciding whether to use voice authentication or when setting up the system for the first time.

In addition, *cognitive forcing* strategies can support people in their decision-making process and have been applied in medical research and research on the overreliance of AI (Bucinca et al., 2021; Croskerry, 2003). They have been described as a "specific debiasing technique that introduces self-monitoring of decisionmaking [sic!]" (Croskerry et al., 2013). We previously adapted cognitive forcing strategies to CAI to make people consider alternatives and reconsider disclosure (Leschanowsky et al., 2022). Depending on the number of possible alternatives, the consideration or generation of alternatives could result in an active choice condition (Table2).

Active choice can help to induce reflection, to avoid mindless acceptance of default options, and to overcome decision avoidance (Keller et al., 2011). Choice architects might have difficulties in coming up with sensible default options as privacy preferences are subjective and heterogeneous. Choice paradigms have been predominantly explored for graphical user interfaces, particularly in the context of cookie consent notices, leaving the need to apply these insights to CAI systems (Habib et al., 2022; Utz et al., 2019). As long as choices are simple, e.g., asking users whether they like to have their data deleted or stored in a certain use case, alternatives can be presented directly to the users and CAI systems can require users to actively choose among them. However, active choice imposes a high cognitive load on individuals and therefore should not be used excessively but applied sensibly (Thaler & Sunstein, 2021). Similar to generating alternatives and cognitive forcing, active choice can be seen as a tool that can effectively support users' decision-making. Other actors in the CAI ecosystem are less likely to benefit from active choice as privacy requirements play a more crucial role than their individual privacy preferences.

As users' privacy decisions may suffer from an underestimation of risks, instructions that make people think of opposite outcomes as initially expected and *prospective hindsight* can counter optimistic privacy choices. By utilizing prospective hindsight, people are asked to imagine their future selves and to experience bad outcomes of their earlier judgements (Mitchell et al., 1989). For example, a conversational AI system might proactively ask users to imagine their future selves 2 years from now and to question why their personal information has been shared with company X and used for profiling. Such a strategy can prevent people from being overly optimistic that privacy breaches will not affect them and therefore trigger the usage of privacy protective strategies. Moreover, prospective hindsight can be triggered by dialog editors that may be used by providers, developers, and integrators to create CAI. This could make them consider the impact of privacy breaches on their reputation, revenue, and employment and help to design CAI with protective strategies in mind. Lastly, owners can similarly benefit by regularly utilising prospective hindsight.

To increase the *accuracy of the judgement*, multiple judgements by others or the same person at different times or with mental focus can be beneficial (Herzog &

Hertwig, 2009; Larrick, 2004). As decisions are based on only a subset of accessible information, subsets can vary once people are asked to rely on different decision strategies, e.g., making one intuitive and one thoughtful decision (Lambe et al., 2016). In the privacy context, the accuracy of judgements could relate to how well users' decisions match their attitudes and values. Consequently, users' regrets and frustration about privacy decisions might decrease while their overall satisfaction might increase. Therefore, users' satisfaction with their privacy decision-making could benefit from offering them the option to reconsider their decision or asking them to decide twice following different instructions—one based on intuitive thinking and one based on in-depth and analytical thinking (Leschanowsky et al., 2022).

Similarly, *planned interruptions* or forced breaks provide another way to introduce reflection and increase the accuracy of judgements. Similar techniques such as diagnostic time-outs or slowing-down decision-making have been successfully applied in the medical field (Lambe et al., 2016). Moreover, Wang et al. (2013) investigated a timer nudge as one of several privacy nudges on social media. The timer nudge would delay the Facebook post and allow users to reflect and possibly cancel their actions. Overall, their timer nudge was perceived positively as it provided the chance to correct typos, post better quality content, or cancel unnecessary posts. In a recent chatbot experiment, we investigated the impact of a timer nudge on users' behavior but found that the additional delay did not significantly impact their decisionmaking (Leschanowsky et al., 2023). While the timer nudge did not negatively affect the usability in our study, conversation designers might be unlikely to apply forced breaks as they generate friction and let the conversation appear less natural. However, planned interruptions can support providers, integrators, and application developers to reflect on their usage of users' personal information and reduce unnecessary permission requests.

Finally, *planning prompts* ask people to specify "when, where, and how" a goal is achieved (Wust & Beck, 2018). These concrete plans help to translate goals into actions and become a commitment which individuals are less likely to break. Therefore, planning prompts provide a simple and effective nudge for goal achievement. Cuadra et al. (2021) explored planning prompts for virtual voice assistants and found that their voice application was perceived as helpful and improved planning behavior. In the privacy context, planning prompts could support both developers as well as users of conversational AI systems. System providers, developers, integrators, and owners could use planning prompts to make specific plans for incorporating privacy into their system and application or to delete unused and old data. Moreover, users can benefit from conversational AI systems that proactively encourage them to make plans for checking their privacy settings.

Limitations A variety of cognitive strategies could support actors in the CAI ecosystem to make better decisions about their privacy. However, based on our assessment, not every cognitive strategy is suitable for all actors. Future research should investigate various cognitive strategies and their suitability to different actors in the CAI system. While only few cognitive strategies have been applied and evaluated in privacy scenarios (Wang et al., 2013), most strategies are yet to be investigated in the context of CAI. Therefore, evaluation measures are needed to test the strategies' effectiveness on people's decision-making. While medical research assesses their effectiveness by evaluating error rates in diagnostic reasoning (Lambe et al., 2016), error rates are not easily accessible in the privacy context. In particular, when investigating cognitive strategies for users of CAI systems who come with highly subjective privacy preferences and attitudes, the optimal outcome of a privacy decision usually remains unknown. We will further discuss the need for a comprehensive evaluation in Section 4.2.3.

Takeaway

Cognitive strategies aim to mitigate bias in human judgements at the time of decision-making and to trigger a more rational thinking process. CAI systems can assist users in generating alternative choices or make them consider alternatives and reconsider decisions by applying cognitive forcing strategies. As long as choices are easily understandable, they should be directly presented through active choice mechanisms. Prospective hindsight and planning prompts are two of the cognitive strategies that can support all actors in the CAI ecosystem to improve their judgements. Instead, planned interruptions might be most suitable to providers, developers, and integrators as they can cause considerable friction to the dialog.

4.2.3 Assisted Decision-Making

Assisted Decision-Making for Debiasing Decision-Making Another way to prevent people from biased decision-making is to replace human judgements with automated or assisted decisions, e.g., by applying linear models or using decision support systems (Larrick, 2004). While these models can incorporate human judgement and subjective ratings, they mostly rely on historical data and objective ratings. However, as mentioned earlier, an individual's privacy preferences are highly subjective and contextdependent (Nissenbaum, 2010). Therefore, the selection of suitable attributes that need to be included in a predictive model is challenging and can again be prone to cognitive biases (Soll et al., 2015).

Assisted Decision-Making in CAI In the privacy context, *privacy assistants* have been investigated which can offer varying levels of automation. While some privacy assistants only inform users and ask them to make decisions, others automatically decide for the users (Colnago et al., 2020) (Table 3).

In addition, *checklists* can be seen as a tool for assisted decision-making as they provide an efficient, systematic, and consistent way of carrying out tasks (Gawande, 2009). They are especially useful in situations of low decision readiness or when certain tasks are likely to be overlooked and left out (Gawande, 2009). In previous work, we showed how checklists can be adapted to privacy in conversational AI systems, e.g., by setting up a privacy checklist and confirming user-specific privacy requirements before installing a new application (Leschanowsky et al., 2022). Moreover, checklists can provide a helpful tool for all actors in the conversational AI ecosystem. They can ensure that app developers follow certain steps that are necessary to protect users' privacy, e.g., checking whether the data asked for is truly relevant. Similarly, providers can benefit from a privacy-related checklist to ensure that data flows on the platform are appropriate and purposeful. For example, the nonprofit organization Open Voice Network has realized the need for privacy checklists and released ethical guidelines that can be interpreted as privacy checklists for voice interfaces (Open Voice Network, 2022, 2023). Lastly, checklists can be distributed to owners to ensure that privacy guidelines are followed when setting up the devices.

Limitations In their study on privacy assistants for IoT, Colnago et al. (2020) found little consensus among users regarding the level of automation and possible control options. Therefore, allowing users to adjust and configure privacy assistants to their needs is highly recommended. Yet, this raises questions on the effectiveness of privacy assistants as a debiasing strategy as it would require additional mechanisms to make users engage with the tool and its control options. Moreover, models that learn peoples' privacy preferences based on historic data can present a privacy threat themselves and need to be implemented in secure and privacy-friendly ways.

Takeaway
Assisted decision-making replaces human judgements altogether and can therefore prevent biased decision-making. Users can benefit from privacy assistants deployed on conversational AI systems, but because of varying preferences, suitable control options are indispensable. In addition, checklists can be considered a tool for assisted decision-making and can support all actors in the ecosystem to make efficient and consistent choices.

4.2.4 Incentives

Incentives for Debiasing Decision-Making Monetary as well as non-financial incentives such as badges or peer pressure have been proven beneficial for people to make better decisions (Acquisti et al., 2018; Lindbeck, 1997). Thereby, incentives can serve as motivators to transition between system 1 and system 2, i.e., between fast and slow thinking, and can be especially useful when undesired choices stem from insufficient attention or a lack of effort (Larrick, 2004). Privacy costs are often difficult to assess as they require an estimate of long-term consequences. Therefore, individuals may have a clear understanding of benefits while the costs remain elusive and hard to grasp (Leschanowsky et al., 2021). Thus, providing the right incentives can make costs understandable and help people in considering long-term consequences. Incentives can be either rewarding, e.g., rewarding individuals for privacy-preserving decisions or considering privacy costs, or punishing, e.g., increasing the costs to choose non-privacy-preserving options or disclosing information about costs and negative consequences of insecure behavior (Lindbeck, 1997). Incentives always present modifications to the environment (see Fig. 2 for their classification into our framework). Yet, while incentives such as organizational measures or regulations focus on influencing decision-making in the future, incentives that are directly rewarding or punishing can also influence peoples' choices at the time of decision-making.

Incentives in CAI System providers, developers, integrators, and owners can profit largely by introducing the right incentives themselves as well as by being exposed to them. For instance, *virtual badges or app reviews* can act as strong incentives for app developers to offer privacy-preserving applications (Acquisti et al., 2018). Similarly, badges and reviews can motivate owners of CAI systems in accommodation places to protect users' privacy and security. Moreover, being accountable and liable for security failures, e.g., having to pay increased fines or to pass additional training, has been shown to create a moral hazard (Acquisti et al., 2018; Anderson, 2001). Being held accountable increases the cost of failure and consequently the effort of making a desired decision (Larrick, 2004; Lerner & Tetlock, 1999) (Table 4).

Herath and Rao (2009) investigated factors that influence employees' intentions to comply with security policies. They found that intrinsic, i.e., perceived effectiveness, and extrinsic motivators, i.e., penalties and social pressure, influence employee behavior. However, while the certainty that possible security breaches are detected positively influenced employees' behavior, the severity of penalties did not. Thus, CAI system providers and integrators should make use of efficient and visible *organizational measures* to detect privacy breaches without needing to severely punish detected breaches. This might add to a positive error culture. Moreover, Herath and Rao (2009) found that social pressure and normative beliefs can strongly impact security behavior. This is particularly interesting from the privacy perspective as privacy breaches are often a result of inappropriate internal information flows rather than security issues. Therefore, system providers and integrators should emphasize privacy practices and expectations throughout the company and communicate to employees the importance of their individual privacy practices.

However, we need to be clear that market forces alone are not sufficient for system providers to push for privacy-preserving solutions and innovations (Stucke & Ezrachi, 2017). Therefore, additional incentives such as *regulations* need to be set by legislators and policymakers to ensure that privacy measures are integrated. Sætra (2020) shows that if privacy is seen as an "aggregate public good," governmental interventions and regulations are beneficial and necessary. In addition to regulations and incentives based on penalties for those who fail to comply, policymakers can include rewarding incentives in their portfolio. Kosseff (2016) refers to "positive cybersecurity law" where companies are encouraged to protect themselves from cybersecurity attacks. While they focus solely on cybersecurity protection, policies like a "safe harbor from data security lawsuits" or tax incentives could be extended towards privacy.

Limitations One limitation of incentives is characterized by the nature of incentives themselves. As most of them focus on enhancing decision-making in the future by introducing new regulations or measures, their adoption takes time and might need to be accentuated by additional guidelines, trainings, and similar interventions (Dalela et al., 2022).

Takeaway
Incentives can be either rewarding or punishing and can be of a financial as well as non-financial nature. Actors designing, implementing, and deploying CAI systems can benefit from virtual badges, reviews, organizational measures, and regulations. These incentives can be either created by themselves or by a higher-level body, e.g., by policymakers.

4.2.5 Defaults

Defaults for Debiasing Decision-Making As humans often stick to default options due to the status quo bias and the difficulty in overcoming inertia, defaults are powerful tools for "choice architects" (Thaler & Sunstein, 2021). It has been shown that defaults can have a significant impact on people's decision-making in fields such as retirement savings, food consumption, and health care (Acquisti et al., 2018; Thaler & Sunstein, 2021). Due to their power and robustness, default options are particularly important to ensure privacy in conversational AI systems.

Defaults in CAI In fact, researchers have argued for implementations of privacy-friendly *default settings* (Table 5), e.g., storage of voice commands and their usage for the system's improvement should be disabled by default (Hernández Acosta & Reinhardt, 2022; Lau et al., 2018). However, current conversational AI systems rarely follow these recommendations and come with varying default settings. For instance, Amazon Alexa default options include an unlimited retention period of voice recordings while Apple's Siri does not retain audio recordings by default (Amazon Inc., 2022a; Apple Inc., 2022).
So far, we have only touched on privacy defaults for users of conversational AI systems, but similar privacy-friendly defaults need to be investigated for system providers, developers, integrators, and owners. For example, Amazon does not share any voice recordings with third-party skill developers, and skills can be configured to request permissions (Amazon Inc., 2022b). However, Lentzsch et al. (2021) found that instead of making permission requests through the API, skills can access users' personal information by asking them directly in a conversation. While making the permission request the default way to access personal information might seem sufficient from a technical point of view, practical implementations prove it wrong. This urges the need for system providers to re-design tools that are used for building CAI and to make privacy-friendly defaults more sticky and less likely to be circumvented by developers. For instance, dialog editors could detect whenever developers ask for personal information in their application and display a prompt that allows them with one click to request the desired information through the API. This would shorten dialogs and has the potential to improve user experience as well as data collection transparency.

Limitations While default options can lead to more privacy-preserving systems, they might not serve all users equally well. Defaults should be deployed for individual and public welfare especially when they are set in place by policymakers. However, problems may arise as defaults that benefit a majority could be suboptimal for some people (Smith et al., 2013). Several studies on CAI systems have shown that while there is some agreement on certain privacy aspects, e.g., implementation of shorter retention periods for voice recordings (Malkin et al., 2019), people's privacy preferences can differ largely (Lau et al., 2018). Therefore, more research is needed on people's privacy preferences in conversational AI to curate sensible default options. In addition, studies in other fields have shown that if easily adjustable controls are offered, people overcome inertia and change defaults if they dislike the outcome (Thaler & Sunstein, 2021). Such controls should be investigated for conversational AI to accompany defaults.

Smith et al. (2013) discuss defaults as "hidden persuadors" and their potential to erode consumers' autonomy. To counter this, they suggest the usage of "smart defaults" which are based on consumer information and adapted to optimally fit a specific consumer. While these smart defaults work well for a variety of contexts, e.g., "Advanced Air Bag System" (Smith et al., 2013), they cause problems in the context of privacy. First, they require access to personal information to adapt to individual preferences, a procedure that comes with varying privacy risks. Second, individuals' valuations of privacy are inconsistent and sensitive to non-normative factors (Acquisti et al., 2013). By trading away privacy for convenience or economic benefits, privacy protection may be led by individual interest rather than social welfare. Therefore, smart defaults (and similar models such as privacy assistants as discussed in Section 4.2.3) can only provide suitable protection once a certain level of privacy is ensured by design. Lastly, defaults can significantly lose impact once companies and consumer interests are not aligned. Based on the example of tracking, Willis (2014) provide an extensive argument why defaults are likely to fail as long as companies can push back and leverage similar biases to make default options more or less sticky. Drawing on information-cost theory, Bar-Gill and Ben-Shahar (2021) show how current legal regulations, such as the GDPR and California Consumer Privacy Act (CCPA), attempt to make privacy-preserving defaults more sticky by asking for explicit consent while at the same time reducing the cost for people to become informed by requiring easily understandable notices. Lowering the information costs can thus result in more users acting upon their attitudes. However, it has yet to be investigated how privacy notices could be made easily understandable in CAI. As they are based on natural language, written notices may need to be translated into dialogues for text- and audio-based interactions.

Takeaway

Nudging with defaults can be used for all actors in the CAI ecosystem by curating sensible defaults for application settings and tools used to design, develop, and deploy these systems. To design acceptable defaults, preferences need to be understood and control options to easily change defaults need to be available. Control options should be available both in dialogs and graphical interfaces to make them easily accessible and more sticky. In addition, information costs and other biases have to be considered as they may influence the effectiveness of defaults.

4.2.6 Nudging with Information and Presentation

Information and Presentation for Debiasing Decision-Making While general education on privacy literacy can support users in overcoming their biases and making more informed decisions, nudges that inform users at the time of decision-making can additionally prove beneficial. Especially, in situations where decision readiness might be low, additional nudges that disclose information warn or remind people can lead to improved decision-making. Shaping information in a way that is intuitive to understand and evokes interest can encourage people to make better decisions (Acquisti et al., 2018; Thaler & Sunstein, 2021).

Nudging with Information and Presentation in CAI *Privacy and security labels* are one option to support people during purchase or download of applications. For instance, Emami-Naeini et al. (2020) developed a two-layered privacy and security label for IoT devices based on expert and user studies. Along the same lines, Johansen et al. (2022) discuss privacy labels and their potential from a multidisciplinary perspective. They also show how privacy labels can not only have an educating effect on users but how these labels can benefit programmers in integrating privacy into their development.

Privacy labels disclose privacy and security information in a more understandable and easily readable format and can support people in their purchasing decisions. However, it might still be difficult to compare privacy labels across multiple products. In their book, Thaler and Sunstein (2021) favor so-called *choice engines* that can help consumers to decide between many alternatives. For instance, travel websites allow users to search among many different options based on their preference selection. Once disclosures about privacy and security attributes are machine-readable, choice engines can allow users to easily compare between varying products and filter for privacy and security options. First attempts to this can be seen by Mozilla (2022) who created a guide to help shopping secure products with the option to choose between varying categories, e.g., smart home and health care applications, and to filter for products where "privacy is and is not included." Moreover, Tamò et al. (2021) propose a right to customization where companies are asked to offer multiple variants with different data processing options and trade-offs between privacy and utility. To support users in their decision-making, choice engines can play a crucial role in comparing applications and products across companies as well as within companies (Table 6).

While presentation nudges for privacy and security have mostly focused on graphical user interfaces (Acquisti et al., 2018; Kitkowska et al., 2020), it is unclear how information and presentation nudges can be applied in conversational AI and in particular voice-enabled systems. For instance, Pearman et al. (2022) refined a consent flow for the US Health Insurance Portability and Accountability Act (HIPAA) authorization in a text-based chatbot. Their iterative process of redesign aimed to make the consent form shorter, clearer, and easier to understand. They found that while their redesigns improved understandability, it was not sufficient to ensure informed consent and recommend *conversational privacy* to tackle shortcomings. Harkous et al. (2016) proposed "Conversational Privacy Bots (PriBots)" which can present privacy policies and enable changing of privacy settings in natural language. Brüggemeier and Lalone (2022) explored conversational privacy in a chatbot by allowing users to control their data or ask for privacy-related information in natural language. They found differences in perception between an offer to delete data and an offer to delete sensitive data. Only the option to delete data was perceived as significantly more private and secure indicating the need to explore possible framing and priming effects and their impact on privacy decision-making. Moreover, while research on anthropomorphism has shown to significantly impact peoples' perceptions of conversational AI systems (Cai et al., 2022; Ha et al., 2021; Ischen et al., 2020), it is unknown how changes to the visual appearance or voice can effect users in their privacy perceptions and behavior. In addition to conversational approaches, other modalities for information and presentation nudges need to be explored for voice-enabled CAI systems and could be developed by providers, integrators, and owners. Yeasmin et al. (2020) investigated modalities for privacy notifications in varying contexts and user preferences. They distinguished between visual and audio notifications and notifications via SMS, email, or app. While user preferences varied depending on the context, a majority preferred audio and visual notifications and notifications via app.

Lastly, *warnings and reminders* can serve as nudges to support developers in integrating security and privacy into their workflow. Thereby, it is important to note that developers are mostly aware of necessary security measures but lack knowledge of privacy practices (Balebako & Cranor, 2014). Nevertheless, in both cases, nudges that provide information can benefit developers and consequently their users. On the security side, security advice integrated into cryptographic APIs has been shown to significantly reduce insecure code (Gorski et al., 2018). On the privacy side, Peddinti et al. (2019) tested a nudge to inform mobile app developers of unnecessary permission requests. Thereby, they included information about permission requests of similar applications to incentivize developers to minimize personal data usage. They found that nudges were effective in reducing permission requests across a broad range of mobile application categories. Moreover, Tahaei et al. (2021) investigated framing nudges on developers with respect to mobile advertising networks. Among other conditions, they presented application developers with a privacy-focused framing that explained the impact of personalized ads on user privacy. Developers exposed to these options were significantly more likely to choose non-personalized ads over personalized ads, and most of them expressed the need to protect users' privacy.

Limitations Previous research has largely focused on nudging strategies for graphical user interfaces (Acquisti et al., 2018; Ioannou et al., 2021; Kitkowska et al., 2020). Yet, CAI systems can come without screens and require adaptable ways of presenting information to users. While research has started investigating new information and presentation nudges (Brüggemeier & Lalone, 2022; Harkous et al., 2016; Pearman et al., 2022; Yeasmin et al., 2020), there remain many open challenges. Future research could investigate the interplay between anthropomorphic features and conversational privacy as well as the influence of context on conversational privacy. Moreover, efforts need to be taken to allow comparability among CAI systems with respect to their privacy and security. While first evaluation frameworks are available to combat unethical design in CAI (Mildner et al., 2022), further research on standardized design guidelines and evaluation measures is required to ensure lawful and ethical design.

Takeaway

While previous attempts on nudging with information and presentation have been successful to influence privacy and security decision-making and can be adapted to CAI systems, CAI poses unique challenges to informing owners and users. Privacy labels and choice engines can be adjusted to fit the context of CAI and help owners and users to make more informed choices. In addition, new information and presentation nudges that resemble CAI's modalities need to be explored. Conversational privacy can leverage CAI's unique capabilities to communicate in natural language to inform owners and users about the system's privacy and security. Lastly, warnings and reminders can nudge all actors in the ecosystem towards privacy-preserving choices.

5 Discussion and Future Work

In this position paper, we make two main contributions about applying debiasing strategies in the context of conversational AI. First, we establish a categorization framework for debiasing strategies based on previous research (Croskerry et al., 2013; Soll et al., 2015) and adapt existing privacy debiasing strategies to the context of CAI (see Fig. 2). Second, we assign those strategies to the relevant stakeholders of the CAI ecosystem as defined by European Data Protection Board (EDPB) (2021). Our proposed debiasing framework can serve as a suitable starting point to further investigate debiasing strategies for CAI but does not come without limitations and future research challenges.

First, we did not include a detailed overview of cognitive biases and heuristics in CAI systems as our focus was primarily on debiasing strategies and their application to CAI. However, due to the human-like nature of CAI systems and the complexity of the CAI ecosystem, CAI-specific biases might arise that should be explored in future work. Moreover, as described in Section 4.2, biases can also be specific to individual actors. Consequently, a comprehensive mapping of biases to actors could inform the design of novel debiasing strategies. Moreover, people or organizations can take on combinations of roles as shown in Section 4.1. Our discussion of debiasing strategies to the actors can help to identify useful strategies for such combined roles.

Second, our discussion has focused on mitigating biases in individual judgements rather than focusing on decision-making in groups. However, as conversational AI systems are often designed, developed, and deployed by a team of engineers and developers, mitigating bias on an individual level might not be sufficient. Importantly, strategies that are suitable for individuals might even introduce new biases on the group level (Kerr & Tindale, 2004). Therefore, future work should explore available approaches for groups to foster unbiased privacy and security decision-making.

Third, we focused on well-known debiasing strategies and their adaptation to CAI. Yet, social sciences and the medical field is especially rich in various debiasing strategies we have not addressed in this position paper, e.g., strategies based on pre-commitement (Ariely & Wertenbroch, 2002; Lambe et al., 2016). Future research could investigate additional debiasing strategies for privacy and security decision-making in CAI. Our proposed framework can thereby help to classify new strategies and set them in relation to existing ones. As all these interventions come with strengths and weaknesses, they should be seen as complementary to each other. Thus, our framework can support the development of holistic solutions by applying combinations of debiasing strategies to support people in their privacy decision-making.

Fourth, as a starting point, we focused on actors defined by the European Data Protection Board (EDPB) (2021). Yet, due to the complexity of the CAI ecosystem, other actors are likely to play a role in the design, implementation, and deployment of the systems. Moreover, we have only slightly touched on the role of policymakers and regulators in incentivizing privacy and security. As policymakers have relied on nudges to assist decision-making in many fields (Thaler & Sunstein, 2021), they form an influential group and should be considered in more detail in future work.

While we aimed at providing an overview of debiasing strategies from different fields, i.e., social science, medical field, privacy, and security, we acknowledge that there are research directions that we have only slightly touched upon or have not considered. For example, we included virtual badges and app reviews in Section 4.2.4—a debiasing strategy that falls into the area of gamification. Gamification focuses on triggering intrinsic motivation through the adoption of game elements and has developed independently from nudging and behavioral economics (De Troyer, 2021). Nevertheless, gamification can support people in their privacy and security decision-making, e.g., by incentivizing privacy and security design or by applying them to educational settings to enhance individuals' privacy literacy. Recent years have seen combined research on gamification and nudging to support sustainability behavior, engagement in mental health applications, or mitigation of cognitive biases (Auf et al., 2021; Dunbar et al., 2014; Luger-Bazinger & HornungPrähauser, 2021). Therefore, future research should explore gamification as a meta-strategy and its effect on debiasing privacy and security decision-making in CAI.

Moreover, we focused on conversational AI systems in general including text-based as well as voice-based systems. Yet, these differences in modality can influence peoples' perceptions (Cho, 2019). Future research could investigate debiasing strategies for various modalities and explore their differences and commonalities. Here, our categorization framework can help to compare debiasing strategies across modalities.

Finally, we want to urge the need for a comprehensive evaluation framework for debiasing strategies for privacy and security decision-making. So far, only a few studies have discussed potential evaluation measures and guidelines for ethical nudge design (Acquisti et al., 2018; Barev et al., 2021; Renaud & Zimmermann, 2018). Yet, they do not provide quantitative measures to evaluate the effectiveness of debiasing strategies. Instead, the privacy field could benefit from drawing on evaluation measures used in the medical field or on studies on cognitive control and rational decision-making (Kahneman, 2011; Lambe et al., 2016; Mushtaq et al., 2011). In more recent work, Habib and Cranor (2022) present an evaluation framework for privacy choice mechanisms. Thereby, they include the aspect of neutrality to evaluate privacy choice mechanisms to address nudging patterns and in particular dark patterns that nudge users away from privacy-protective options. Yet, such a framework does not take into account bright patterns and nudging strategies towards privacy-preserving behavior. Moreover, only a few have focused on evaluating debiasing strategies in light of current legal regulations and their legitimacy (Barev et al., 2021, 2022). Therefore, interdisciplinary research is necessary to pave the way for effective and legitimate debiasing strategies for privacy decision-making in CAI.

Funding Open Access funding enabled and organized by Projekt DEAL. Our work is funded by the German Federal Ministry for Economic Affairs and Energy as part of their AI innovation initiative (funding code 01MK20011A).

Data Availability We do not analyze or generate any datasets, because our work proceeds within a theoretical approach.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Abdi, N., Ramokapane, K. M., & Such, J. M. (2019). More than smart speakers: Security and privacy perceptions of smart home personal assistants. *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019),*451–466. USENIX Association.

Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L.F., Komanduri, S., ... & Wilson, S. (2018). Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Computing Surveys, 50*, 1–41.

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *American Association for the Advancement of Science, 347*, 509–514.

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2020). Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. 30, 736–758. Wiley Online Library. https://doi.org/10.1002/jcpy.1191

Acquisti, A., John, L. K., & Loewenstein, G. (2013). What is privacy worth? *The Journal of Legal Studies, 42*, 249–274. University of Chicago Press Chicago, IL. https://doi.org/10.1086/67175

Alashoor, T., Al-Maidani, N., & Al-Jabri, I. (2018). The privacy calculus under positive and negative mood states.

Alepis, E., & Patsakis, C. (2017). Monkey Says, Monkey Does: Security and Privacy on Voice Assistants. *IEEE Access, 5*, 17841–178510. https://doi.org/10.1109/ACCESS.2017.2747626

Almuhimedi, H., Schaub, F., Sadeh, N., Adjerid, I., Acquisti, A., Gluck, J., ... & Agarwal, Y. (2015). Your Location has been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 787–796. Association for Computing Machinery. https://doi.org/10.1145/2702123.2702210

Amazon Inc. (2019). *Whitepaper: Alexa privacy and data handling overview (20191220v2).* Retrieved 24 October 2022, from https://aws.amazon.com/de/alexaforbusiness/resources/?a4b-whats-new.sort-by=item.additionalFields.postDateTime&a4b-whats-new.sort-order=desc

Amazon Inc. (2022a). *Alexa history: See, hear and delete your voice recordings.* Retrieved 24 October 2022 from https://www.amazon.com/alexa-history-delete-voice-recordings/b?ie=UTF8&node=21137870011

Amazon Inc. (2022b). Configure permissions for customer information in your skill. Retrieved 24 October 2022 from: https://developer.amazon.com/en-US/docs/alexa/custom-skills/configure-permissions-for-customer-information-in-your-skill.html.

Anderson, R. (2001). Why information security is hard-an economic perspective. *Seventeenth annual computer security applications conference*, 358–365.

Apple Inc. (2022c). *Privacy*. Retrieved 24 October 2022 from: https://wwwapple.com/privacy/features/.

Ariely, D., & Wertenbroch, K. (2002). Procrastination, deadlines, and performance: Self-control by precommitment (Vol. 13, pp. 219–224). SAGE Publications Sage CA: Los Angeles, CA. https://doi.org/10.1111/1467-9280.00441

Auf, H., Dagman, J., Renström, S., & Chaplin, J. (2021). Gamification and nudging techniques for improving user engagement in mental health and well-being apps. *Proceedings of the Design Society, 1*, 1647– 1656. Cambridge University Press. https://doi.org/10.1017/pds.2021426

Bach, A. K. P., Nørgaard, T. M., Brok, J. C., & Van Berkel, N. (2023, April). If i had all the time in the world: ophthalmologists' perceptions of anchoring bias mitigation in clinical ai support. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–14. Hamburg Germany: ACM. https://doi.org/10.1145/35445483581513

Balebako, R., & Cranor, L. (2014). Improving App Privacy: Nudging App Developers to Protect User Privacy. *Security & Privacy, IEEE, 12*, 55–58. https://doi.org/10.1109/MSP.2014.70

Barev, T.J., Dickhaut, E., Schomberg, S., Janson, A., Schöbel, S., Grote, T., ... & Leinmeister, J.M. (2022). *Handlungsbroschüre. Systematisches Design digitaler Privacy Nudges*. Kassel University Press.

Barev, T. J., Schöbel, S., Janson, A., & Leimeister, J. M. (2021). Delen–a process model for the systematic development of legitimate digital nudges. *16th International Conference on Design Science Research in Information Systems and Technology, DESRIST 2021,* 299–312.

Bar-Gill, O., & Ben-Shahar, O. (2021). Rethinking Nudge: An Information Costs Theory of Default Rules. *University of Chicago Law Review*, *88*, 531–604.

Beattie, H., Watkins, L., Robinson, W. H., Rubin, A., & Watkins, S. (2022). Measuring and mitigating bias in ai-chatbots. *2022 Ieee International Conference On Assured Autonomy (icaa)*, 117–123. https://doi.org/10.1109/ICAA52185.2022.00023

Bispham, M., Zard, C., Sattar, S., Ferrer-Aran, X., Suarez-Tangil, G., & Such, J. (2022). Leakage of Sensitive Information to Third-Party Voice Applications. *4th Conference on Conversational User Interfaces*, 1–4. ACM. https://doi.org/10.1145/3543829.3544520

Bispham, M. K., van Rensburg, A. J., Agrafiotis, I., & Goldsmith, M. (2020). Black-Box Attacks via the Speech Interface Using Linguistically Crafted Input. P. Mori, S. Furnell, & O. Camp (Eds.), Information Systems Security and Privacy (Vol. 1221, pp. 93–120). Springer International Publishing. https://doi.org/10.1007/978-3-030-49443-85

Broussard, J. D., & Wulfert, E. (2019). Debiasing strategies for problem gambling: *Using decision science to inform clinical interventions, 6*, 175–182. Springer. https://doi.org/10.1007/s40429-019-00263-1

Brüggemeier, B., & Lalone, P. (2022). Perceptions and reactions to conversational privacy initiated by a conversational user interface. Computer Speech & Language (Vol. 71, p. 101269). https://doi.org/10.1016/j.csl2021.101269

Bucinca, Z., Malaya, M. B., Gajos, & K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AIassisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 1–21. https://doi.org/10.1145/3449287

Cai, R., Cain, L. N., & Jeon, H. (2022). Customers' perceptions of hotel AIenabled voice assistants: does brand matter? *International Journal of Contemporary Hospitality Management*, *34*, 2807–2831. https://doi.org/10.1108/IJCHM-10-2021-1313

Cavoukian, A. (2009). *Privacy by design: The 7 foundational principles* (Vol. 5).

Cho, E. (2019). Hey google, can i ask you something in private? Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (p. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300488

Colesky, M., Caiza, J. C., Del Alamo, J. M., Hoepman, J.-H., & Martín, Y.-S. (2018). A system of privacy patterns for user control. *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 1150–1156. ACM. https://doi.org/10.1145/3167132.3167257

Colnago, J., Feng, Y., Palanivel, T., Pearman, S., Ung, M., Acquisti, A., ... & Sadeh, N. (2020). Informing the Design of a Personalized Privacy Assistant for the Internet of Things. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376389

Croskerry, P. (2003). Cognitive forcing strategies in clinical decisionmaking. *Annals of Emergency Medicine, 41*, 110–120.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 2: impediments to and strategies for change. BMJ quality & safety (Vol. 22, pp. ii65–ii72). BMJ Publishing Group Ltd. https://doi.org/10.1136/bmjqs-2012-001713

Cuadra, A., Bankole, O., & Sobolev, M. (2021). Planning Habit: Daily Planning Prompts with Alexa. R. Ali, B. Lugrin, & F. Charles (Eds.), *Persuasive Technology* (Vol. 12684, pp. 73–87). Springer International Publishing. https://doi.org/10.1007/978-3-030-79460-6 7

Dalela, A., Giallorenzo, S., Kulyk, O., Mauro, J., & Paja, E. (2022). A study on security and privacy practices in danish companies. *Usable Security and Privacy (USEC) Symposium 2022.*

De Troyer, O. (2021). Gamification, persuasive techniques, and nudging: What is the impact on the user experience? *RoCHI*, 1–3. https://doi.org/10.37789/rochi.2021.1.1.1

Dinev, T., McConnell, A. R., & Smith, H. J. (2015). Research commentary—informing privacy research through information systems, psychology, and behavioral economics: *thinking outside the "apco" box, 26*, 639–655. INFORMS. https://doi.org/10.1287/isre.2015.0600

Dunbar, N. E., Miller, C. H., Adame, B. J., Elizondo, J., Wilson, S. N., Lane, B. L., ... & Zhang, J. (2014, August). Implicit and explicit training in the mitigation of cognitive bias through the use of a serious game. *Computers in Human Behavior*, *37*, 307–318. https://doi.org/10.1016/j.chb.2014.04.053

Edu, J., Ferrer-Aran, X., Such, J., & Suarez-Tangil, G. (2022). Measuring alexa skill privacy practices across three years. *Proceedings of the ACM Web Conference 2022,* 670–680. ACM. https://doi.org/10.1145/3485447.3512289

Emami-Naeini, P., Agarwal, Y., Faith Cranor, L., & Hibshi, H. (2020). Ask the Experts: What Should Be on an IoT Privacy and Security Label? *2020 IEEE Symposium on Security and Privacy (SP)*, 447–464. IEEE. https://doi.org/10.1109/SP40000.2020.00043

European Data Protection Board (EDPB). (2021). *Guidelines 02/2021 on Virtual Voice Assistants (Version 2.0) https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-022021-virtual-voice-assistants_en* Online accessed 22 November 2022

Fernandes, D., Lynch, J., & Netemeyer, R. (2014). Financial Literacy, Financial Education, and Downstream Financial Behaviors. *Management Science*. https://doi.org/10.1287/mnsc.2013.1849

Garg, R., & Sengupta, S. (2019). ”When you can do it, why can’t I?”: Racial and Socioeconomic Differences in Family Technology Use and Non-Use. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 1–22). https://doi.org/10.1145/3359165

Garg, R., & Sengupta, S. (2020). Conversational Technologies for In-home Learning: Using Co-Design to Understand Children’s and Parents’ Perspectives. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–13). ACM. https://doi.org/10.1145/3313831.3376631

Gawande, A. (2009). *The checklist manifesto: how to get things right*. Metropolitan Books.

Gorski, P. L., Iacono, L. L., Wermke, D., Stransky, C., Möller, S., Acar, Y., & Fahl, S. (2018). Developers deserve security warnings, too: On the effect of integrated security advice on cryptographic API misuse. *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, 265– 281. USENIX Association.

Ha, Q.-A., Chen, J. V., Uy, H. U., & Capistrano, E. P. (2021). Exploring the Privacy Concerns in Using Intelligent Virtual Assistants under Perspectives of Information Sensitivity and Anthropomorphism. *International Journal of Human–Computer Interaction*, *37*, 512–527. https://doi.org/10.1080/10447318.2020.1834728

Habib, H., & Cranor, L. F. (2022). Evaluating the usability of privacy choice mechanisms. *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, 273–289.

Habib, H., Li, M., Young, E., & Cranor, L. (2022). “okay, whatever”: An evaluation of cookie consent interfaces. *Proceedings of the 2022 chi conference on human factors in computing systems* (pp. 1–27).

Harkous, H., Fawaz, K., Shin, K. G., & Aberer, K. (2016). PriBots: Conversational privacy with chatbots. *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016).* USENIX Association.

Hatamian, M. (2020). Engineering privacy in smartphone apps: a technical guideline catalog for app developers. *IEEE Access*, *8*, 35429–35445. (Conference Name: IEEE Access) https://doi.org/10.1109/ACCESS.2020.2974911

Herath, T., & Rao, H. (2009). Encouraging information security behaviors in organizations: Role of penalties, pressures and perceived effectiveness. *Decision Support Systems*, *47*, 154–165. https://doi.org/10.1016/j.dss.2009.02.005

Hernández Acosta, L., & Reinhardt, D. (2022). A survey on privacy issues and solutions for voice-controlled digital assistants. *Pervasive and Mobile Computing, 80*, 101523. https://doi.org/10.1016/j.pmcj2021.101523

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. (Vol. 20, pp. 231–237). SAGE Publications Sage CA: Los Angeles, CA. https://doi.org/10.1111/j.1467-9280.2009.02271.x

Ioannou, A., Tussyadiah, I., Miller, G., Li, S., & Weick, M. (2021). Privacy nudges for disclosure of personal information: A systematic literature review and meta-analysis. *Plos One*, *16*, e0256822. Public Library of Science. https://doi.org/10.1371/journal.pone.0256822

Ischen, C., Araujo, T., Voorveld, H., van Noort, G., & Smit, E. (2020). Privacy Concerns in Chatbot Interactions. A. Følstad et al. (Eds.), *Chatbot Research and Design*, *11970*, 34–48. Springer International Publishing. https://doi.org/10.1007/978-3-030-39540-73

Johansen, J., Pedersen, T., Fischer-Hübner, S., Johansen, C., Schneider, G., Roosendaal, A., ... & Noll, J. (2022). A multidisciplinary definition of privacy labels. *Information & Computer Security*, *30*, 452–469. https://doi.org/10.1108/ICS-06-2021-0080

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar Straus and Giroux.

Keller, P. A., Harlam, B., Loewenstein, G., & Volpp, K. G. (2011). Enhanced active choice: A new method to motivate behavior change. *Journal of Consumer Psychology, 21*, 376–383). Elsevier. https://doi.org/10.1016/j.jcps.2011.06.003

Kerr, N. L., & Tindale, R. S. (2004, February). Group Performance and Decision Making. *Annual Review of Psychology*, *55*, 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009

Kitkowska, A., Shulman, Y., Martucci, L. A., & Wästlund, E. (2020). Psychological effects and their role in online privacy interactions: a review. *IEEE Access, 8*, 21236–21260. IEEE. https://doi.org/10.1109/ACCESS.2020.2969562

Kocielnik, R., Avrahami, D., Marlow, J., Lu, D., & Hsieh, G. (2018). Designing for workplace reflection: a chat and voice-based conversational agent. *Proceedings of the 2018 Designing Interactive Systems Conference*, 881–894. ACM. https://doi.org/10.1145/3196709.3196784

Kosseff, J. (2016). Positive cybersecurity law: creating a consistent and incentive-based system symposium: cyberwars: navigating responsibilities for the public and private sector. *Chapman Law Review*, *19*, 401–420.

Kröger, J. L., Gellrich, L., Pape, S., Brause, S. R., Ullrich, S. (2022). Personal information inference from voice recordings: User awareness and privacy concerns. *Proceedings on Privacy Enhancing Technologies*, *2022*, 6–27. https://doi.org/10.2478/popets-2022-0002

Lambe, K. A., O'Reilly, G., Kelly, B. D., & Curristan, S. (2016). Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Quality & Safety*, *25*, 808–820. https://doi.org/10.1136/bmjqs-2015-004417

Larrick, R. P. (2004). Debiasing. *Blackwell handbook of judgment and decision making* (pp. 316–338). Blackwell Publishing Ltd Malden, MA, USA. https://doi.org/10.1002/9780470752937.ch16

Lau, J., Zimmerman, B., & Schaub, F. (2018). Alexa, are you listening?: privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, *2*, 1–31. https://doi.org/10.1145/3274371

Lentzsch, C., Shah, S. J., Andow, B., Degeling, M., Das, A., & Enck, W. (2021). Hey Alexa, is this skill safe?: taking a closer look at the alexa skill ecosystem. *Proceedings 2021 Network and Distributed System Security Symposium.* Internet Society. https://doi.org/10.14722/ndss.2021.23111

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*, 255. American Psychological Association. https://doi.org/10.1037/0033-2909.125.2.255

Leschanowsky, A., Brüggemeier, B., & Peters, N. (2021). *Design Implications for human-machine interactions from a qualitative pilot study on privacy* (pp. 76–79). Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication. https://doi.org/10.21437/SPSC.2021-16

Leschanowsky, A., Popp, B., & Peters, N. (2022). *Adapting debiasing strategies for conversational AI.* Zagreb, Croatia, 74.

Leschanowsky, A., Popp, B., & Peters, N. (2023). *Privacy strategies for conversational AI and their influence on users' perceptions and decision-making*. Proc. 2023 European Symposium on Usable Security (EuroUSEC). https://doi.org/10.1145/3617072.3617106

Liao, S., Wilson, C., Cheng, L., Hu, H., & Deng, H. (2020). Measuring the effectiveness of privacy policies for voice assistant applications. *Annual Computer Security Applications Conference*, 856–869. ACM. https://doi.org/10.1145/3427228.3427250

Lindbeck, A. (1997). Incentives and social norms in household behavior. *The American Economic Review, 87*, 370–377. JSTOR.

Ludolph, R., & Schulz, P. J. (2018). Debiasing health-related judgments and decision making: a systematic review. *Medical Decision Making*, *38*, 3–13. Sage Publications Sage CA: Los Angeles, CA. https://doi.org/10.1177/0272989x17716672

Luger-Bazinger, C., & Hornung-Prähauser, V. (2021). Innovation for sustainable cities: The effects of nudging and gamification methods on urban mobility and sustainability behaviour. *GI Forum 2021*, *9*, 251–258. Verlag der Osterreichischen Akademie der Wissenschaften. https://doi.org/10.1553/giscience202102s251

Malkin, N., Deatrick, J., Tong, A., Wijesekera, P., Egelman, S., & Wagner, D. (2019). Privacy attitudes of smart speaker users. *Proceedings On Privacy Enhancing Technologies, 2019.* https://doi.org/10.2478/popets-2019-0068

Mann, M., Wilson, M., & Warren, I. (2022). Smart parenting? The internet of things, children's privacy, and data justice. *The International Journal of Children's Rights*, *30,* 204–231. https://doi.org/10.1163/15718182-30010008

Maroufkhani, P., Asadi, S., Ghobakhloo, M., Jannesari, M. T., & Ismail, W. K .W. (2022). How do interactive voice assistants build brands' loyalty? *Technological Forecasting and Social Change*, *183*, 121870. https://doi.org/10.1016/j.techfore.2022.121870

Masur, P. K. (2019). Situational Privacy and Self-Disclosure. *Springer International Publishing*. https://doi.org/10.1007/978-3-319-78884-5

Masur, P. K., Teutsch, D., Dienlin, T., & Trepte, S. (2017). Online Privat he its kompetenz und deren Bedeutung fu¨r demokratische Gesellschaften. *Forschungsjournal Soziale Bewegungen*, *30*, 180–189. https://doi.org/10.1515/fjsb-2017-0039

McTear, M. (2021). *Conversational ai: dialogue systems, conversational agents, and chatbots*. Springer International Publishing. https://doi.org/10.1007/978-3-031-02176-3

Mildner, T., Doyle, P., Savino, G.-L., Malaka, R. (2022). Rules of engagement: levelling up to combat unethical cui design. *4th Conference on Conversational User Interfaces,* 1–5. ACM. https://doi.org/10.1145/3543829.3544528

Mitchell, D. J., Edward Russo, J., & Pennington, N. (1989). Back to the future: Temporal perspective in the explanation of events. *Journal of Behavioral Decision Making*, 2, 25–38. Wiley Online Library. https://doi.org/10.1002/bdm.3960020103

Mozilla. (2022). *Be smart. shop safe.* Retrieved 24 October 2022 from https://foundation.mozilla.org/en/privacynotincluded/.

Muntwiler, C. (2021). Debiasing management decisions: Overcoming the practice/theory gap within the managerial decision process. *Proceedings of Take 2021 Conference,* 123.

Murad, C., Munteanu, C., R. Cowan, B., & Clark, L. (2021). Finding a new voice: transitioning designers from gui to vui design. *CUI 2021 3rd Conference on Conversational User Interfaces*, 1–12. ACM. https://doi.org/10.1145/3469595.3469617

Murtarelli, G., Gregory, A., & Romenti, S. (2021). A conversation-based perspective for shaping ethical human–machine interactions: the particular challenge of chatbots. *Journal of Business Research*, *129*, 927–935. https://doi.org/10.1016/j.jbusres.2020.09.018

Mushtaq, F., Bland, A. R., Schaefer, A. (2011, October). Uncertainty and cognitive control. *Frontiers in Psychology*, *2*, 249. https://doi.org/10.3389/fpsyg.2011.00249

Neal, T., & Brodsky, S. L. (2016). Forensic psychologists' perceptions of bias and potential correction strategies in forensic mental health evaluations. *Psychology, Public Policy, and Law*, *22*, 58. American Psychological Association. https://doi.org/10.1037/law0000077

Nissenbaum, H. F. (2010). *Privacy in context: technology, policy, and the integrity of social life*. Stanford Law Books.

Nutt, P. C. (2004). Expanding the search for alternatives during strategic decision-making. *The Academy of Management Executive (1993–2005), 18*, 13–28.

Open Voice Network. (2022). *Privacy and security work group meeting.* OVON Privacy and Security Work Group Meeting October, 25th 2022.

Open Voice Network. (2023). *Ethical guidelines for voice experiences Version 2.0.*https://openvoicenetwork.org/docs/ethical-guidelines-for-voice-experiences*. Online accessed 21 August 2023.*

OpenAI. (2023). *Gpt-4 technical report.* arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Orphanou, K., Otterbacher, J., Kleanthous, S., Batsuren, K., Giunchiglia, F., Bogina, V., ... & Kuflik, T. (2022). Mitigating bias in algorithmic systems—a fish-eye view. *Acm Computing Surveys*, *55*, 1–37. ACM New York, NY. https://doi.org/10.1145/3527152

Pal, D., Arpnikanondt, C., Razzaque, M. A., Funilkul, S. (2020). To Trust or not-trust: privacy issues with voice assistants. *IT Professional*, *22*, 46–53. https://doi.org/10.1109/MITP.2019.2958914

Park, Y. J. (2013). Digital literacy and privacy behavior online. *Communication Research*, *40*, 215–236. https://doi.org/10.1177/0093650211418338

Pearman, S., Young, E., & Cranor, L.F. (2022). User-friendly yet rarely read: a case study on the redesign of an online HIPAA authorization. *Proceedings on Privacy Enhancing Technologies*, *2022*, 558–581. https://doi.org/10.56553/popets-2022-0086

Peddinti, S. T., Bilogrevic, I., Taft, N., Pelikan, M., Erlingsson, U., Anthonysamy, P., & Hogben, G. (2019). Reducing permission requests in mobile apps. *Proceedings of the Internet Measurement Conference*, 259–266. ACM. https://doi.org/10.1145/3355369.3355584

Rao, S., Resendez, V., El Ali, A., & Cesar, P. (2022). Ethical self-disclosing voice user interfaces for delivery of news. *4th Conference on Conversational User Interfaces,* 1–4. ACM. https://doi.org/10.1145/3543829.3544532

Renaud, K., & Zimmermann, V. (2018). Ethical guidelines for nudging in information security & privacy. *International Journal of Human-Computer Studies*, *120*, 22–35. Elsevier. https://doi.org/10.1016/j.ijhcs.2018.05.011

Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2022). Voice in human–agent interaction: a survey. *ACM Computing Surveys, 54*, 1–43. https://doi.org/10.1145/3386867

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1–20. https://doi.org/10.1146/annurev.ps.41.020190000245

Sin, J., Munteanu, C., Ramanand, N., & Tan, Y. R. (2021). VUi influencers: how the media portrays voice user interfaces for older adults. *CUI 2021 - 3rd Conference on Conversational User Interfaces*, 1–13. ACM. https://doi.org/10.1145/3469595.3469603

Singh, R. (2019). Profiling Humans from their Voice. *Springer Singapore*. https://doi.org/10.1007/978-981-13-8403-5

Smith, N. C., Goldstein, D. G., & Johnson, E. J. (2013). Choice without awareness: ethical and policy implications of defaults. *Journal of Public Policy & Marketing, 32*, 159–172. https://doi.org/10.1509/jppm.10.114

Soll, J. B., Milkman, K. L., Payne, J. W. (2015). A user's guide to debiasing. G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (pp. 924–951). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118468333.ch33

Stucke, M. E., & Ezrachi, A. (2017). How digital assistants can harm our economy, privacy, and democracy. *Berkeley Technology Law Journal*, 32, 1239–1300. JSTOR.

Subramaniam, M., Kumar, P., Morehouse, S., Liao, Y., & Vitak, J. (2019). Leveraging funds of knowledge to manage privacy practices in families. *Proceedings of the Association for Information Science and Technology*, *56*, 245–254. https://doi.org/10.1002/pra2.67

Sundar, S. S., & Kim, J. (2019). Machine heuristic: when we trust computers more than humans with our personal information. *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–9. Association for Computing Machinery. https://doi.org/10.1145/3290605.3300768

Sætra, H. S. (2020). Privacy as an aggregate public good. *Technology in Society*, *63*, 101422. https://doi.org/10.1016/j.techsoc.2020.101422

Tahaei, M., Frik, A., & Vaniea, K. (2021). Deciding on personalized ads: nudging developers about user privacy. *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, 573–596.

Tamò Larrieux, A., Zihlmann, Z., García, K., & Mayer, S. (2021). Right to customization: conceptualizing the right to repair for informational privacy. Springer.

Terzopoulos, G., & Satratzemi, M. (2020). Voice Assistants and Smart Speakers in Everyday Life and in Education. *Informatics in Education*, *19*, 473–490. https://doi.org/10.15388/infedu.2020.21

Thaler, R., & Sunstein, C. (2021). *Nudge: The final edition*. Penguin Publishing Group.

UC Berkeley School of Information. (2019). *Privacy patterns*. Retrieved 24 October 2022 from https://privacypatterns.org/

United Nations. (1948). *Universal declaration of human rights*. UN General Assembly.

Utz, C., Degeling, M., Fahl, S., Schaub, F., & Holz, T. (2019). (un)informed consent: Studying gdpr consent notices in the field. *Proceedings of the 2019 acm sigsac conference on computer and communications security* (p. 973–990). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3319535.3354212

van Mil, J., & Quintais, J. P. (2022). A matter of (Joint) control? Virtual assistants and the general data protection regulation. *Computer Law & Security Review*, *45*, 105689. https://doi.org/10.1016/j.clsr.2022105689

Wang, Y., Leon, P. G., Scott, K., Chen, X., Acquisti, A., & Cranor, L. F. (2013). Privacy nudges for social media: An exploratory facebook study. *Proceedings of the 22nd International Conference on World Wide Web,* 763–770. Association for Computing Machinery. https://doi.org/10.1145/2487788.2488038

Welch, C. F., Pérez-Rosas, V., Kummerfeld, J. K., & Mihalcea, R. (2019). Look who's talking: Inferring speaker attributes from personal longitudinal dialog. *Conference on Intelligent Text Processing and Computational Linguistics*.

Willis, L. E. (2014). Why not privacy by default. *Berkeley Technology Law Journal*, *29*, 61–134.

Wilson, C. G., Bond, C. E., & Shipley, T. F. (2019). How can geologic decisionmaking under uncertainty be improved? *Solid Earth*, *10*, 1469–1488. Copernicus GmbH. https://doi.org/10.5194/se1014692019

Wüst, K., & Beck, H. (2018). "i thought i did much better"—overconfidence in university exams. Decision sciences journal of innovative education (Vol. 16, pp. 310–333). Wiley Online Library. https://doi.org/10.1111/dsji.12165

Yeasmin, F., Das, S., & Backstrom, T. (2020). Privacy analysis of voice user interfaces. *Conference of Open Innovations Association, FRUCT*, 6.

Zargham, N., Reicherts, L., Bonfert, M., Völkel, S.T., Sch¨oning, J., Malaka, R., & Rogers, Y. (2022). Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. *Proceedings of the 4th Conference On Conversational User Interfaces,* 1–14.