Ethical Design of Conversational Agents: Towards Principles for a Value-Sensitive Design

Thiemo Wambsganss
Universität St.Gallen

Anne Höch
Universität Kassel

Naim Zierau
Universität St.Gallen

Matthias Söllne
Universität Kassel

Ethical Design of Conversational Agents: Towards Principles for a Value-Sensitive Design

Thiemo Wambsganss[1], Anne Höch[2], Naim Zierau[1], and Matthias Söllner[2]

[1] University of St.Gallen (HSG), Institute of Information Management, St.Gallen, Switzerland
{thiemo.wambsganss, naim.zierau}@unisg.ch
[2] University of Kassel, Information Systems and Systems Engineering, Kassel, Germany
annehoech@web.de, soellner@uni-kassel.de

Abstract. Conversational Agents (CAs) have become a new paradigm for human-computer interaction. Despite the potential benefits, there are ethical challenges to the widespread use of these agents that may inhibit their use for individual and social goals. However, besides a multitude of behavioral and design-oriented studies on CAs, a distinct ethical perspective falls rather short in the current literature. In this paper, we present the first steps of our design science research project on principles for a value-sensitive design of CAs. Based on theoretical insights from 87 papers and eleven user interviews, we propose preliminary requirements and design principles for a value-sensitive design of CAs. Moreover, we evaluate the preliminary principles with an expert-based evaluation. The evaluation confirms that an ethical approach for design CAs might be promising for certain scenarios.

Keywords: Ethics in IS, Value-Sensitive Design, Conversational Agents, Design Science Research

1 Introduction

Driven by technological advances in *Artificial Intelligence* (AI) especially in the area of *Natural Language Processing* (NLP), many organizations strive to leverage the potential of Conversational Agents (CAs) for improving *human-computer interaction* (HCI) [1]. CAs such as *Amazon's Alexa, Google's Assistant, and Apple's Siri* are software programs that engage with users through natural language [2]. CAs promise to dramatically enhance user experience by enabling personalization, around the clock availability, and immediate response times [3]. The popularity of these interfaces has been steadily growing over the past few years [3]. Thus, a plethora of positive user outcomes have been recorded, such as engagement [4], trust [5, 6], rapport, and ease of use, in several domains, such as education [7–9], healthcare [10] and customer service [11, 12]. Despite the potential benefits of these agents, there are ethical problems that arise from the use of many contemporary CAs. First, the appearance and behavior of CAs are susceptible to design biases such that certain stereotypes are reinforced and strengthened. For instance, [13] found that most agents are embodied with feminine characteristics, as these are supposed to improve the attitude towards the agents, but also solidify specific gender roles. Second, the knowledge base and respective Machine

16th International Conference on Wirtschaftsinformatik,
March 2021, Essen, Germany

Learning (ML) models are susceptible to bias, resulting in systematic errors that may create unfair outcomes. For example, they can lead to inaccurate predictions for specific subgroups or may carry the implicit values of programmers and organizations [14]. Moreover, these agents operate with some level of autonomy, resulting in increased opaqueness that highlights questions of accountability and transparency [15]. For instance, [16] has shown that users are more likely to choose financial portfolios that exceed their risk profiles when using a CA compared to non-conversation robo advisors, which may serve as an example of how these agents can be used to manipulate customers. Finally, as CAs operate on user data and may, in fact, be used to collect enormous amounts of (sensitive) data, user privacy becomes an even more important issue [17]. In sum, while CAs have the potential to fundamentally improve user outcomes, developers and providers may need to increasingly follow ethical considerations in the design of these agents to ensure the well-being of their users [18].

However, as the proliferation of CAs has been driven mostly by monetary goals (e.g., [20]), it remains doubtful whether the design of these agents takes ethical concerns sufficiently into account and could rightfully give rise to the skepticism of many users. This sentiment is also reflected in CA research to date, as most authors did not follow a distinct normative approach in deriving design guidelines but rather descriptively analyze user interactions with these agents, which does not allow to draw direct conclusions about the ethical design of these agents (e.g., [21]). In fact, the importance of ethical perspectives on design research on novel IS artifacts has been discussed by IS scholars long before. For example, [22] stated that ethical considerations in the field of IS design should receive a more prominent role. Following this, [23] suggested to include ethical guidelines in the design research of IS artifacts and proposed six ethical principles informing design science research. [24] followed by discussing the philosophical responsibility of IS research. Recently, IS researchers have identified the novelty of AI-based CAs as IS artifacts and called for further work to investigate ethical designs with principles and guidelines for CAs [1, 25]. Also, in practice, value-sensitive design plays a prominent role, i.e., large technology providers of CAs such as Google[1] and Microsoft[2] have recently released ethical guidelines on the design of these AI systems. Moreover, intergovernmental organizations such as the *Organization for Economic Co-operation and Development* (OECD) or the *Group of Twenty* (G20) drive the societal debate by releasing principles for the ethical design of AI systems such as CAs (OECD[3] in May 2019, G20[4] in June 2019). The intergovernmental guidelines and current literature strongly motivate the need for a value-sensitive design of AI-driven IS such as CAs. However, they provide a more conceptual framing with rather general categories for AI-based IS artifacts [23]. Current literature falls short to provide meaningful and evaluated design principles (e.g., according to [26]) to help IS designers and practitioners to 1) instantiate value-sensitive CAs and 2) evaluate currently instantiated CAs from a value-sensitive design perspective based on these principles.

[1] https://ai.google/principles/
[2] https://www.microsoft.com/en-us/ai/responsible-ai
[3] https://www.oecd.org/going-digital/ai/principles/
[4] https://dig.watch/updates/g20-digital-economy-ministers-endorse-ai-principles

Following the *AI principles of the OECD*, we therefore aim to contribute to the field of value-sensitive design of CA by answering the following research question **(RQ)**:

***RQ**: What are relevant design principles that foster a value-sensitive design of Conversational Agents?*

To answer the stated research question, we overall follow a design science-research approach (DSR). As stated above, there is a lack of concrete design knowledge for the ethical design of CAs. Thus, we intend to iteratively derive and evaluate design knowledge on the baseline of existing normative design recommendations (i.e., OECD AI principles), while focusing on social response theory [27, 28] as a guiding theoretical lens to inform concrete artifact design [29]. Users experience those agents as increasingly human-like, which is why social response theory may represent a new *"foundation for understanding and designing humane anthropomorphic agents"* ([25], p.1). In sum, we follow a value-sensitive design approach that allows us to translate ethical requirements or imperatives (i.e., OECD AI principles) identified into actionable design guidelines [31]. To the best of our knowledge, there is no study that rigorously derives requirements from both scientific literature and potential users to derive design principles for value-sensitive CAs following intergovernmental guidelines, such as the OECD AI principles. With a value-sensitive CA, we implicate a dialogue-based system that incorporates human and ethical values into its core design and implementation process, e.g., when designing the interaction or when training ML models.

In this paper, we present the preliminary design principles and an expert-based evaluation of those principles according to [32]. Our results suggest that a value-sensitive design of CAs might be a promising approach for different user interaction scenarios, e.g., where privacy and transparency play an important role. With a further evaluation of these design principles, they might serve as a foundation informing CA designers towards an ethical design. In the following, we will first introduce the reader to the necessary theoretical background. Afterwards, we present our methodological approach for creating design knowledge following the three cycle view of [33]. Finally, our preliminary requirements and design principles are presented and evaluated by experts, followed by an outline of the subsequent steps and the expected implications once our research is completed.

2 Theoretical Background

2.1 Value-Sensitive Design of Conversational Agents

Recent advances in NLP and ML bear the opportunity to design new forms of HCI for IS with conversational interfaces, also called Conversational Agents (CAs). CAs are software programs that are designed to communicate with users through natural language interaction interfaces [2]. In today's world, conversational interfaces, such as *Amazon's Alexa*, *Google's Assistant,* or *Apple's Siri*, are ubiquitous, with their

popularity steadily growing over the past few years [34]. They are implemented in various areas, such as customer service [12, 35], counseling [36], collaboration [37] or education [7, 38]. Recently, an overwhelming amount of research emerged in different disciplines that investigated the effect of different design elements and configurations unique to these agents on various forms of user perceptions, such as trust or social presence (e.g., [12, 21]). However, the application of AI usually comes with disadvantages, such as lower transparency, loss of control, and lack of trust by human users [39]. As [25] claim, *current ethical design perspectives fall mostly short of a practical application of design principles for the interaction design of CAs*. Value-based design is a theoretically grounded approach for a technological design that integrates human values in a principled and understandable way during the whole design process [31]. Ethics can be seen as a foundation of value-sensitive or value-based design [22]. Nevertheless, literature strongly motivates the need for a value-sensitive design of AI-driven IS such as CAs but provides a rather conceptual framing with general categories for IS artifacts (such as [23]). Literature only poorly provides meaningful and evaluated design principles to help IS designers and practitioners to instantiate value-sensitive CAs. Thus, we aim to contribute to research by investigating design principles based on the *OECD AI guideline* and therefore follow the recent call for future work by several IS scholars to "*[build] a cumulative body of prescriptive [design] knowledge on methods for the engineering of humane anthropomorphic agents [CAs] as well as generic design principles guiding the design of humane anthropomorphic agents"* ([25], p.14).

The widespread application of AI-based IS has been driving a recent discussion of their values and ethics (i.e., [25]). Earlier studies already focused on different but singular aspects of ethical values, for example, privacy [40], prevention of bias [41] or trust [42]. However, there is a lack of holistic and actionable design knowledge that supports value-sensitive development of novel AI-based IS such as CAs. Besides, several intergovernmental organizations such as the OECD or the G20 have stated principles for AI. The guidelines are complemented by a discourse in the academic literature (e.g., [1, 15, 29]). The OECD collected five ethical principles for AI-based systems by 50 experts from 20 governments as well as leaders from the business, labor, civil society, academic and science communities:

"1) AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being. 2) AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society. 3) There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them. 4) AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed. 5) Organizations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles."

Nevertheless, the OECD AI principles are not operationalized in such a way that allow designers to easily translate them into concrete design features (e.g., according

to [26]). The principles of the OECD were composed to provide a general perspective on the value-sensitive design of AI-based systems and therefore fall short to provide meaningful and evaluated design principles to help IS designers and practitioners to 1) instantiate value-sensitive CAs (e.g., according to [26]) and 2) evaluate existing interaction designs. Therefore, we aim to address this literature gap and investigate, derive and evaluate design principles for value-sensitive CA design.

2.2 Social Response Theory as a Lens for Value-Sensitive IS Design

Our design approach is anchored in social response theory. According to this theory, humans tend to respond socially to IS that displays characteristics similar to humans (e.g., to animals or technologies) [44]. Behavioral clues and social signals from computers, such as interacting with others, using natural language, or playing social roles, subconsciously trigger responses from humans, no matter how rudimentary those clues or signals are [27, 28]. Following the "*Computers are Social Actors"* (CASA) paradigm, existing research has examined different social clues and their influence on HCI (e.g., [21]). However, a value-sensitive and ethical perspective on designing CAs has been poorly considered in the literature, thus inhibiting the development of truly social actors (i.e., agents that act on moral principles [30]). Accountability, transparency, or trust have been proven to play a major role in trustworthy social relationships but are only minorly engrained in the interaction design of CAs (e.g., [45]). Thus, we follow the value-sensitive model of the humanness of CAs [25] by investigating principles for an ethical CA. We aim to contribute to better user acceptance, experience, and user-centered design according to social response theory [27, 28].

3 Research Methodology

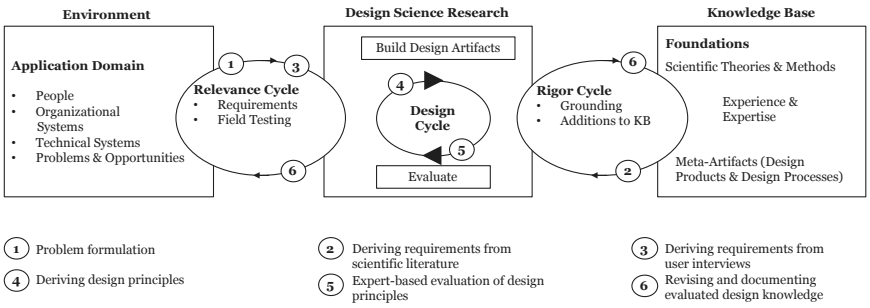

Figure 1. Overview of our design science research approach

To answer our research question, we follow a DSR approach [33]. We decided to follow this methodology, as it allows us to solve a set of practical problems and to contribute to the existing body of knowledge by designing and evaluating new design knowledge based on a sound understanding of the current knowledge base and user

perceptions of a new technological phenomenon [46]. Moreover, this allows us to give a "voice" to the users – a key aspect of value-sensitive design. Figure 1 shows the steps that are carried out.

We focus on translating the OECD AI principles two to five into actionable design principles according to [26]. The first principle depicts rather a meta-principle that encompasses all of the following, which is why we do not include it in this research project that focuses on actionable design knowledge. Overall, our research project aims to contribute to research with a *nascent design theory* that gives explicit prescriptions for a value-sensitive and thus a more ethical design of CAs [47]. We followed a theory-driven design approach by grounding our research on social response theory [27, 28]. The *first step* of the DSR cycle includes the problem formulation. The relevance of the practical problem was therefore described in the introduction of this work. In the *second step*, we derived a set of requirements in the form of *literature issues* (LIs) from the current state of scientific literature for the design of value-sensitive CAs according to the OECD AI principles. Therefore, we conduct a systematic literature review in the fields of Human-Computer Interaction (HCI) and Information Systems (IS) design. Next, we conducted eleven semi-structured interviews with students and professionals using the expert interview method by [48] to capture requirements from users for ethical CAs. Based on the interviews, we gathered *user stories* (USs) as user requirements for the design of a value-sensitive CA. In the *fourth step*, we derived preliminary *design principles* (DPs) addressing the LIs and USs from the prior steps using the structure suggested by [26]. We argue that a CA (and possibly also other AI-based IS) that instantiates our DPs should increase the perceived humanness and thus improve overall user experience and interaction. Our principles should provide designers of CAs with ethical considerations based on implicit values derived from literature and expert interviews. Thus, we aim to enable designers to design more ethical CAs, ultimately increasing the well-being of its users. Accordingly, in *step five*, we perform an expert-based evaluation of our preliminary design principles based on the evaluation framework proposed by [32]. We interview experts from academia and industry to quantitatively and qualitatively evaluate the relevance, robustness, and usefulness of our principles according to the OECD guidelines for designers from the fields of social science, psychology and IS design. At the end of the study, we contribute to research with evaluated design knowledge on how to design value-sensitive CAs based on the OECD AI principles. Overall, we hope to contribute with our findings to *a nascent design theory* [47] for value-sensitive design of CAs.

4 Deriving Design Knowledge

In this section, we will describe and discuss how we derived the preliminary DPs. The problem formulation (step one) described in the introduction serves as the foundation for the derivation of the requirements from literature and users.

4.1 Step 2: Deriving Requirements from Scientific Literature

To derive requirements from scientific literature, a systematic literature search was conducted using the methodological approaches of [49] and [50]. We initially focused our research on studies that demonstrate the successful implementation of a value-sensitive design for IS artifacts. In order to do this, publications on design, ethics and design science of IS and CAs were identified by a systematic search in different search engines and databases, such as *Google Scholar, EBSCO, JSTOR, ACM, AIS Library*. We used the following keywords to find potential hits for our literature review: *"Value-sensitive Design", "User experience", "Chatbot", "Conversational Agent", "Design", "Design Science Research", "Design Artifact", "Ethical AI", "AI Principles", "AI Guidelines"* AND *"Transparency", "Fairness", "Explainability", "Understandability", "Accountability", "Robustness", "Security", "Safety, "Privacy"*. Initially, we received several thousand hits based on these search terms. Therefore, we screened the titles and abstracts of the publications. Our goal was to identify papers that deal with ethical aspects of CAs. Thus, we only included literature that contributes to a kind of ethical perspective on the design of CAs according to the OECD AI principles. We excluded papers that explicitly did not deal with an ethical perspective when deriving design knowledge of CAs (i.e., papers focusing on sales-driven dependent variables). On this basis, we selected 87 papers for more intensive analysis. We have summarized similar topics of these contributions as literature issues (LIs) and formed 15 clusters from them to derive a concept matrix according to [51]. Those topics represent integral design issues that were addressed to increase individual and or social good when using those agents. We allocated those issues to the individual OECD principles, which served as scaffolding divisions for the organization of those issues. The LIs are aggregated and illustrated in Table 1 with exemplary papers.

Table 1. Aggregated LIs for a value-sensitive design of CAs with exemplary papers

Dimension*	#	Literature Issues (LIs)
Human-centered values & fairness	LI1	Prevention of bias or discrimination (e.g., [41])
	LI2	Accessibility & Design (e.g., [52])
	LI3	Compliance with human rights & democratic values (e.g., [53])
	LI4	Beneficence (e. g., [53])
Transparency & explainability	LI5	Transparency (e.g., [54]) & Explainability (e.g., [55])
	LI6	Trust (e.g., [42])
	LI7	Traceability (e.g., [56])
	LI8	Communication (e.g., [57])
Robustness, security & safety	LI9	Non-maleficence (e.g., [53])
	LI10	Privacy (e.g., [40])
	LI11	Resilience (e.g., [58])
	LI12	Reliability (e.g., [42])
Accountability	LI13	Auditability (e.g., [59])
	LI14	Reporting (e.g., [60])
	LI15	Responsibility (e.g., [61])

* according to the OECD AI principles two to five

4.2 Step 3: Deriving Requirements from User Interviews

Based on the derived LIs, we conducted eleven semi-structured interviews according to [48]. The interview guideline consists of 29 questions and each interview lasted around *28 to 59 minutes* (*mean = 40.99 minutes*). The interviewees were all potential users of a value-sensitive CA and all had used a CA before in different scenarios. Therefore, we followed a literal replications logic. Therefore, we chose participants with different insights based on their background (i.e., different demographics). In order to gain impressions resulting from different user groups, a heterogeneous group of users was interviewed, such as students and professionals. The participants were asked about the following topics: experience with CAs, perception of values and ethics in CAs, requirements for a value-sensitive CA (e.g., functionalities, design), requirements for a CA that aims to follow the OECD principles, such as fairness, transparency, robustness, accountability.

The interviewees were in *mean = 32.91 years* old *(SD = 12.06*). Five participants were students of business administration, one of economics, one of teaching profession, and one a student of nutrition science. Three interviewees were practitioners in different sectors (medical, police, and business), four were male, seven were female. After a more precise transcription, the interviews were evaluated using qualitative content analysis. The interviews were coded, and abstract categories were formed. The coding was performed using open coding to form a uniform coding system during evaluation [48]. Based on these results, we gathered *120 user stories* (USs) as user requirements following [62]. We aggregated the most common user stories, which resulted in 19 USs for value-sensitive CAs (illustrated in Table 2).

Table 2. Aggregated user stories for a value-sensitive design of CAs based on [62]

#	User Stories (USs)
US1	As a CA user, I would like to always be treated equally regarding the outcome resulting from collected but not necessarily context-relevant data (e.g., gender, race).
US2	As a CA user, I think it would be helpful if the CA was accessible and available in different language or age groups so that everyone has the same access to benefits/risks.
US3	As a CA user, I think that communication and design should suit different requirements and needs, e.g., older people need more assistance than younger ones, so the system has to be reactive.
US4	As a CA user, impartiality and equality of opportunities and respectful interaction are key for perceiving a CA as fair.
US5	As a CA user, I would like the interaction and communication with the CA to be easy and intuitive.
US6	As a CA user, I wish that the process follows certain structures and is always be understandable.
US7	As a CA user, it would be convenient if the interaction was like human-human interaction in terms of empathy and flexibility.
US8	As a CA user, I expect that the focus is on solving my issues/problems and ensuring this through inquiries and confirmations or exit strategies if necessary.
US9	As a CA user, it would be helpful to know that only context-relevant data is collected to help me and therewith prevent any sort of bias.
US10	As a CA user, it´s important that the system is aware of the potential risk of hacking or theft.
US11	As a CA user, enlightenment and commitment to privacy and data protection rules is inevitable, e.g., regular reports would give me a good feeling/trust in the system.
US12	As a CA user, I would like to have a feedback function and human contact option always available (e.g., as an exit strategy).
US13	As a CA user, in case of an attack, I expect to be directly informed about the attack and what is advised to do, e.g., deleting or changing passwords or that the system is in self-destruction mode and deletes all personal data.
US14	As a CA user, I expect that my data is protected against any kind of abuse.

US15	As a CA user, I would like to use a CA that embeds control mechanisms through independent third parties to make the system credible.
US16	As a CA user, I would like to use a CA that is regularly controlled through independent control organs/institutions for continuous monitoring and improvements.
US17	As a CA user, I would like to use a CA that regularly controls both technical control and human control mechanisms, e.g., to control if intended actions are happening.
US18	As a CA user, I would like to use a CA that reports every action step and provides access to information.
US19	As a CA user, I would like to receive detailed feedback in case something relevant is affecting my data.

4.3 Step 4: Deriving Preliminary Design Principles

As illustrated, we have identified 15 LIs and 19 USs as requirements for a value-sensitive design of CAs. Based on these findings, we derived 14 preliminary DPs for a value-sensitive CA that aim to address OECD AI principles two to five. The design principles (and the LIs as well as the USs the particular DP is derived from) are depicted in Table 3. Our DPs were formulated based on the analysis of current issues related to value-sensitive design, design of CAs and requirements of users based on social response theory [27, 28]. We argue that a CA (and possibly other AI-based ISs) that instantiates our DPs increase the perceived humanness of the CA, for example, through more trustworthy design elements and thus improve the overall user experience and interaction. For example, a value-sensitive CA that employs a mechanism to avoid data bias in the training's data and is instantiated in different languages for different cultural and age backgrounds should be perceived as fairer and human-centered, and thus the interaction with the CA should result in a better user experience.

Table 3. Preliminary Design Principles according to [63]

Dimension	#	Design Principles (DPs)	LI	US
Human-centered values & fairness	**DP1**	For designers to establish a human-centered CA, which is perceived as fair by users, employ a working step that ensures data collection fulfills the minimalism and general data collection regulation to ensure that the user does not feel treated unfairly because of not context-relevant information.	LI4	US1
	DP2	For designers to implement fairness in CA, employ a mechanism that checks the training data for representativeness and bias to make the user feel confident while using the CA and sharing information.	LI 1,3	US1
	DP3	For designers to build a human-centered CA, employ a chat indicator in the design that signals compliance with democratic as well as moral and ethical values to enhance the users' perceived fairness.	LI4	US4
	DP4	For designers to implement a fair and human-centered CA, ensure widespread accessibility and usability to allow users from different language and age backgrounds to easily interact with the CA.	LI2	US2, 3
Transparency & explainability	**DP5**	For designers to enhance the perceived transparency and trustworthiness of a CA, employ an indicator (e.g., some sort of certificate) showing that the CA is compliant with national and international laws and standards to allow the user to perceive the rightful design.	LI6, 8	US5, 6
	DP6	For designers to establish a transparent CA, consider feedback cycles and traceable structures to allow the user to understand internal processes and outcome generation and thus enhance understanding.	LI5, 7,8	US8
	DP7	For designers to employ transparent CAs, integrate an indicator/avatar that educates the user about data collection procedures to allow the user to feel involved and well advised.	LI8	US9
	DP8	For designers to establish transparent and understandable CAs for users, develop a professional wording, flexible (exit strategies, keyword independent solving) and empathetically communicating avatar that creates a convenient and pleasurable user experience.	LI8	US5, 6,7
Robustness	**DP9**	For designers to design robust and secure CAs for users, employ an indicator that signals the user protection of sensitive data and implements safety and	LI9, 10	US11

		security standards/laws to allow the user to feel protected against any type of harm or abuse.		
	DP 10	For designers to establish robust CAs for users, employ regular security checks and well-elaborated risk management strategies that ensure data and privacy security and therewith enhance overall resilience.	LI 11	US10 ,12-14
	DP 11	For designers to develop robust CAs, employ some sort of official certificates in the design that allow the user to strengthen perceived reliability and safety through serious commitment.	LI 12	US11
Accountability	**DP 12**	For designers to establish accountable CAs employ a mechanism that demonstrates independent audit or control organs are regularly revising the CA to ensure compliance with given laws and standards and signals the user trustworthiness for the CA.	LI 13,15	US16
	DP 13	For designers to design accountable CAs, employ an indicator that makes internal reporting strategies and guidelines available for the user and allow for further information and therewith enhance perceived responsibility.	LI 14	US15 , 18
	DP 14	For designers to design accountable CAs, employ logging and tracking mechanisms to establish clear structures that can easily be retraced and understood by users in order to allow for the correct functioning and clear communication towards the user.	LI 14	US17 , 19

4.4 Step 5: Expert-Based Evaluation of Design Principles

In the next step, we aimed to evaluate our preliminary DPs with experts from different domains, such as IS, HCI, and psychology. Our primary goal was to both qualitatively and quantitatively evaluate if the design principles would be of use from the perspective of the experts and if they are robust and important for the design of ethical CAs. Therefore, we performed expert interviews following the criteria of [32]. The interview questionnaire consisted of 46 items and was composed of three parts. We started with an introduction about research on CAs and ethical design to provide a basis for a common understanding. In the second part, we sequentially showed the interviewees our preliminary DPs for a value-sensitive design of CAs and asked questions about their relevance, usability, and robustness. (e.g., "*How important/useful/robust is this DP for an ethical design of CAs and why?"*). We quantitatively captured their impression on a 5-point Likert scale from "fully disagree" to "fully agree". Moreover, we documented their qualitative justification of the answer for each DP. The questionnaire closed with a creative task, where we asked the experts to derive concrete design features on how the DP could be instantiated. We aimed to further evaluate our DPs by analyzing if the experts could deduct a specific design feature from the principle. Moreover, by doing so, we received further design knowledge about potential design instantiations. We provided an empty CA box, where the participants were asked to draw/sketch a design feature or write down their ideas in design statements. In total we interviewed ten experts - eight were researchers, while two were practitioners. The mean age was mean = 28.20 (SD = 8.53), seven were female, three were male. In average the interviews lasted mean = 66.8 minutes (SD = 15.33). The documented results were a) qualitatively evaluated by calculating the mean and standard derivation (SD) for each DP and evaluation dimension (relevance, usefulness, and robustness) and b) qualitatively analyzed by performing a cluster analysis of the provided answers. The quantitative results of our interviews are displayed in Table 4 and the qualitative cluster results along with exemplary quotes are displayed in Table 5.

Table 4. Quantitative results from the expert-based evaluation based on the three evaluation dimensions (relevance, usefulness, and robustness) for each value-sensitive design principle

	Relevance of DP		Usefulness of DP		Robustness of DP	
DP	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
1	4.70	0.46	4.7	0.46	4.0	0.77
2	4.60	0.49	4.4	0.49	4.1	0.94
3	4.4	0.92	4.3	0.90	4.0	0.77
4	4.9	0.30	4.8	0.40	4.7	0.46
5	4.5	0.81	4.6	0.66	4.4	0.80
6	3.8	0.98	4.1	0.70	3.9	0.83
7	4.5	0.67	4.5	0.50	4.1	0.94
8	4.7	0.46	4.8	0.40	4.2	0.87
9	4.5	0.67	4.6	0.66	4.3	1.00
10	4.7	0.64	4.6	0.66	4.7	0.64
11	3.8	1.54	3.9	1.58	3.6	1.50
12	4.6	0.49	4.6	0.49	4.3	1.00
13	4.0	0.77	4.2	0.60	4.3	0.46
14	4.0	1.18	3.3	1.15	3.5	1.29

Our evaluation confirmed that all DPs are mostly positively perceived by the experts in terms of relevance, robustness, and usefulness. The mean values for the DPs are promising when comparing the results to the midpoints of the scale. The relevance of all design principles is better than the neutral value of 3, and all fourteen DPs have normalized values greater than 0.7 (greater absolute values than 3.5), which indicates a high relevance. Regarding the usefulness, only DP14 is evaluated with a mean value lower than 4, which can be explained by the fact that tracing and logging user activities seem to generally be seen sceptical by potential users, which highlights a particularly sensitive area to users that could be meaningfully addressed by value-sensitive design activities. This is reflected in an exemplary expert comment "*Users don´t want their actions and habits to be traced in such detail; they would feel supervised in an uncomfortable way".* Twelve DPs are regarded as highly useful with higher normalized values greater than 4.0 (except for DP 6 and DP 11). Regarding the robustness of the DPs, eleven out of fourteen DPs received higher mean values than 4.0. Only DP11 is regarded as less robust by the experts with a mean = 3.6. The SD for DP11 with regards to robustness is quite high (SD = 1.50) indicating that there is a disappearance between the experts judging the DP as not very robust or as very robust. In their qualitative comments, some interviewees elaborated that they perceive certificates and signaling as important and a way to demonstrate compliance with certain values or procedures. Others formed a completely different image, stating that certificates can also be deceptive and should be seen critically and cautiously, as they are not only advantageous. Eleven DPs are judged as highly robust with greater mean values than 4.0 except for DP 6, DP 11, and DP 14.

Table 5. Clustered qualitative results from the expert-based evaluations by representative examples

Group	Quotes
On fairness (DP1-4)	*"Mitigation bias and ensuring representativeness is of great importance because it enhances perceived fairness."* *"It is harmful if the user has to be concerned about being judged because of criteria (gender, age, religion) that need not be relevant for the outcome."*
On transparency and trust (DP5-8)	*"Certificates support perceived transparency and trust in CAs as a signaling effect, (but be also aware of weaknesses)."* *"Engage in understandable structures and consistent regulations to help users follow the outcome process and reassure with feedback questions that intended goals are reached."* *"I think providing information concerning different topics (e.g., data collection/use, internal risk strategies) is crucial for transparency and trust."* *"I think clear and honest communication about data collection and usage is of great importance. If this is not part of the communication process, the user could feel unsure and is not be likely to trust the system."* *"From my experience users value adequate language and questions as key to perceive a CA as capable and trustworthy."*
On robustness, safety, and security (DP9-11)	*"Sensitive data of users have to stay private to make users feel more confident"* *"To enhance trust and robustness, implement regular safety checks and reduce error tolerance"* *"Especially data security and compliance with the DSGVO should be basic elements of CAs."*
On accountability (DP12-14)	"*An accountable CA should be controlled by independent institutions to ensure regular improvement and compliance with legal norms and standards."* *"Users like to be ensured that internal and external structures/organs prevent misuse or harmful action."*

As described above, we also included qualitative questions in our questionnaire to receive the participants' opinions about the DPs and reasons for their quantitative judgements. The general attitude about most DPs was very positive. Especially the principles for trust and transparency were highlighted often by the interviewees. However, the DPs on accountability sometimes seem to be not clear enough. The clustered results are displayed in Table 5 along with the OECD principles.

Moreover, the experts revealed some interesting ideas and concepts for implementing the DPs as instantiated CAs, e.g., for usability and trust, *"Insert an EXIT-Button to stop interaction or to switch to human."*, or for visualization, *"Insert control units for the size of writing or the sound, in general, be adaptive to different user needs."*, and for different user groups "*Use of symbols instead of texts to makes interaction easier and more intuitive for elderly users*".

5 Discussion and Conclusion

Besides a multitude of behavioral and design-oriented studies on CAs, a distinct ethical perspective on CA design falls rather short in current literature. Therefore, in

our paper, we presented first insights into how to deduct actionable design knowledge for the design of value-sensitive CAs based on contemporary ethical frameworks for AI design. We document 15 literature issues based on 87 papers and 19 user stories based on eleven user interviews on how to design a value-sensitive CA following the OECD AI principles. Moreover, we derived and evaluated 14 design principles that address them. Our results show interesting findings for the design of conversational interfaces and possibly other AI-based IS.

We, therefore, contribute to the design of CAs based on a value-sensitive approach to ensure an ethical perspective on this emerging technology. We provide researchers and practitioners with requirements and design principles for the design of their own CA to help them to ensure their user manipulations are built based on an ethical grounding. Especially with further advances of NLP and ML (e.g., [64]) and newly available data sets for specific domain-related tasks (e.g., argumentation annotated corpora for argumentation skill learning [65, 66]), design knowledge for a value-sensitive perspective on CAs might encourage designers and research towards a more ethical design of these novel ISs. This might help providers of CAs to communicate to the user how a value-sensitive design approach has been followed based on our principles. Overall, we aim to contribute to a *nascent design theory* [47] for the class of value-sensitive IS artifacts. We systematically deduced design knowledge as documented in Table 1, 2, and 3. Due to the systematic procedure, we aimed at generating a satisfying design contribution [67]. We believe with further empirical evaluation and instantiation of our generated design knowledge; we contribute to a nascent design theory in IS (e.g., such as [9]). We, therefore, hope to encourage designers to focus more on an ethical design of conversational IS.

However, our research also comes with limitations. Since our objective was to derive practical design principles to help designers, we derived the requirements from a certain IS and HCI perspective. Different literature streams or different interviewees (e.g., interviews only from ethics) might lead to different results on different granularity levels. Moreover, we abstracted and derived certain design principles to provide a holistic design perspective of the OECD principles. Therefore, a certain abstraction level was chosen. The question that remains for the individual domain and class of CAs is how to instantiate the design principles as design features for their specific use case. Therefore, we call for future work to provide empirical insights into the effects of specific principles and instantiated design features on human perception.

References

1. Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., Söllner, M.: AI-Based Digital Assistants. Bus. Inf. Syst. Eng. 61, 535–544 (2019). https://doi.org/10.1007/s12599-019-00600-8.
2. Shawar, B.A., Atwell, E.S.: Using corpora in machine-learning chatbot systems. Int. J. Corpus Linguist. 10, 489–516 (2005). https://doi.org/10.1075/ijcl.10.4.06sha.
3. De Keyser, A., Köcher, S., Alkire (née Nasr), L., Verbeeck, C., Kandampully, J.: Frontline Service Technology infusion: conceptual archetypes and future research directions. J. Serv.

Manag. 30, 156–183 (2019). https://doi.org/10.1108/JOSM-03-2018-0082.
4. Winkler, R., Hobert, S., Salovaara, A., Söllner, M., Leimeister, J.M.: Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In: Conference on Human Factors in Computing Systems - Proceedings (2020). https://doi.org/10.1145/3313831.3376781.
5. Adam, M., Wessel, M., Benlian, A.: AI-based chatbots in customer service and their effects on user compliance. Electron. Mark. (2020). https://doi.org/10.1007/s12525-020-00414-7.
6. Zierau, N., Engel, C., Söllner, M., Leimeister, J.M.: Trust in Smart Personal Assistants: A Systematic Literature Review and Development of a Research Agenda. In: 15th International Conference on Wirtschaftsinformatik (WI 2020) (2020).
7. Winkler, R., Söllner, M.: Unleashing the Potential of Chatbots in Education : A State-Of-The-Art Analysis . In : Academy of Management. Meet. Annu. Chicago, A O M. (2018).
8. Wambsganss, T., Winkler, R., Schmid, P., Söllner, M.: Unleashing the Potential of Conversational Agents for Course Evaluations: Empirical Insights from a Comparison with Web Surveys. In: Twenty-Eighth European Conference on Information Systems (ECIS2020). pp. 1–18. , Marrakesh, Morocco (2020).
9. Wambsganss, T., Söllner, M., Leimeister, J.M.: Design and Evaluation of an Adaptive Dialog-Based Tutoring System for Argumentation Skills. In: International Conference on Information Systems (ICIS). , Hyderabad, India (2020).
10. Laumer, S., Maier, C., Gubler, F.T.: Chatbot Acceptance in Healthcare: Explaining User Adoption of Conversational Agents for Disease Diagnosis. Twenty-Seventh Eur. Conf. Inf. Syst. (ECIS2019), Stock. Sweden. 0–18 (2019).
11. Følstad, A., Brandtzaeg, P.B.: Users' experiences with chatbots: findings from a questionnaire study. Qual. User Exp. 5, 1–14 (2020). https://doi.org/10.1007/s41233-020-00033-2.
12. Zierau, N., Wambsganss, T., Janson, A., Schöbel, S., Leimeister, J.M.: The Anatomy of User Experience with Conversational Agents : A Taxonomy and Propositions of Service Clues. Icis 2020. 1–17 (2020).
13. Feine, J., Gnewuch, U., Morana, S., Maedche, A.: Gender Bias in Chatbot Design. In: CONVERSATIONS 2019: Chatbot Research and Design. pp. 79–93. Springer (2020). https://doi.org/10.1007/978-3-030-39540-7_6.
14. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., Jennings, N.R., Kamar, E., Kloumann, I.M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D.C., Pentland, A. 'Sandy,' Roberts, M.E., Shariff, A., Tenenbaum, J.B., Wellman, M.: Machine behaviour. Nature. 568, 477–486 (2019). https://doi.org/10.1038/s41586-019-1138-y.
15. Pfeuffer, N., Benlian, A., Gimpel, H., Hinz, O.: Anthropomorphic Information Systems. Bus. Inf. Syst. Eng. 61, 523–533 (2019). https://doi.org/10.1007/s12599-019-00599-y.
16. Hildebrand, C., Bergner, A.: Conversational robo advisors as surrogates of trust: onboarding experience, firm perception, and consumer financial decision making. J. Acad. Mark. Sci. (2020). https://doi.org/10.1007/s11747-020-00753-z.
17. Roßnagel, A.: Smarte Persönliche Assistenten gestalten. Datenschutz und Datensicherheit - DuD. 44, 565–566 (2020). https://doi.org/10.1007/s11623-020-1324-y.
18. Følstad, A., Brandtzaeg, P.B., Feltwell, T., Law, E.L.C., Tscheligi, M., Luger, E.A.: Chatbots for social good. Conf. Hum. Factors Comput. Syst. - Proc. 2018-April, (2018).

https://doi.org/10.1145/3170427.3185372.
19. Fuckner, M., Barthes, J.P., Scalabrin, E.E.: Using a personal assistant for exploiting service interfaces. In: Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design. pp. 89–94 (2014). https://doi.org/10.1109/CSCWD.2014.6846822.
20. Reddy, T.: Chatbots for customer service will help businesses save $8 billion per year, https://www.ibm.com/blogs/watson/2017/05/chatbots-customer-service-will-help-businesses-save-8-billion-per-year/, last accessed 2020/05/01.
21. Feine, J., Gnewuch, U., Morana, S., Maedche, A.: A Taxonomy of Social Cues for Conversational Agents. Int. J. Hum. Comput. Stud. 132, 138–161 (2019). https://doi.org/10.1016/j.ijhcs.2019.07.009.
22. Mingers, J., Walsham, G.: Toward ethical information systems: The contribution of discourse ethics. MIS Q. Manag. Inf. Syst. 34, 855–870 (2010). https://doi.org/10.2307/25750707.
23. Myers, M.D., Venable, J.R.: A set of ethical principles for design science research in information systems. Inf. Manag. 51, 801–809 (2014). https://doi.org/10.1016/j.im.2014.01.002.
24. Hassan, N.R., Mingers, J., Stahl, B.: Philosophy and information systems: where are we and where should we go?, https://www.tandfonline.com/action/journalInformation?journalCode=tjis20, (2018). https://doi.org/10.1080/0960085X.2018.1470776.
25. Gimpel, H., Bayer, S., André, E., Benke, I., Benlian, A., Cummins, N., Hinz, O., Kersting, K., Maedche, A., Riemann, J., Schuller, B., Weber, K.: Humane Anthropomorphic Agents : the Quest for the Outcome Measure. In: AIS SIGPrag 2019 pre-ICIS workshop "Values and Ethics in the Digital Age" (2019).
26. Gregor, S., Chandra Kruse, L., Seidel, S.: The Anatomy of a Design Principle. J. Assoc. Inf. Syst. Forthcomin, (2020).
27. Nass, C., Steuer, J., Tauber, E.R.: Computers are social actors. In: Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94. pp. 72–78. ACM Press, New York, New York, USA (1994). https://doi.org/10.1145/191666.191703.
28. Nass, C., Moon, Y.: Machines and Mindlessness: Social Responses to Computers. J. Soc. Issues. 56, 81–103 (2000). https://doi.org/10.1111/0022-4537.00153.
29. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. Des. Sci. IS Res. MIS Q. 28, 75 (2004).
30. Mädche, A.: Humane Anthropomorphic Agents : The Quest for the Outcome Measure. In: Pre-ICIS Workshop 2019 "Values and Ethics in the Digital Age." pp. 1–18 (2019).
31. Friedman, B., Kahn Jr., P.H., Borning, A.: Value Sensitive Design and Information Systems. Human-Computer Interact. Manag. Inf. Syst. Found. 1–27 (2006). https://doi.org/10.1145/242485.242493.
32. Venable, J., Pries-Heje, J., Baskerville, R.: FEDS: A Framework for Evaluation in Design Science Research. Eur. J. Inf. Syst. 25, 77–89 (2016). https://doi.org/10.1057/ejis.2014.36.
33. Hevner, A.R.: A three cycle view of design science research. Scand. J. Inf. Syst. 1–6 (2007).
34. Krassmann, A.L., Paz, F.J., Silveira, C., Tarouco, L.M.R., Bercht, M.: Conversational Agents in Distance Education: Comparing Mood States with Students' Perception. Creat. Educ. 09, 1726–1742 (2018). https://doi.org/10.4236/ce.2018.911126.

35. Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., Luo, J., Akkiraju, R.: Touch your heart: A tone-aware chatbot for customer care on social media. Conf. Hum. Factors Comput. Syst. - Proc. 2018-April, (2018). https://doi.org/10.1145/3173574.3173989.
36. Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., Armour, C., McTear, M.: Towards a chatbot for digital counselling. HCI 2017 Digit. Make Believe - Proc. 31st Int. BCS Hum. Comput. Interact. Conf. HCI 2017. 2017-July, 1–7 (2017). https://doi.org/10.14236/ewic/HCI2017.24.
37. Elshan, E., Ebel, P.: Let's Team Up: Designing Conversational Agents as Teammates. Int. Conf. Inf. Syst. (2020).
38. Wambsganss, T., Winkler, R., Söllner, M., Leimeister, J.M.: A Conversational Agent to Improve Response Quality in Course Evaluations. In: ACM CHI Conference on Human Factors in Computing Systems (2020).
39. Mädche, A.: Humane Anthropomorphic Agents : The Quest for the Outcome Measure. Pre-ICIS Work. 2019 "Values Ethics Digit. Age." 1–18 (2019).
40. Rosen, J.: Why Privacy Matters. Wilson Q. 24 (Autumn, 32–38 (2000). https://doi.org/10.1145/1378727.
41. Veale, M., Binns, R.: Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data Soc. 4, (2017). https://doi.org/10.1177/2053951717743530.
42. Gefen, D., Karahanna, E., Straub, D.W.: Trust and tam in online shopping: AN integrated model. MIS Q. Manag. Inf. Syst. 27, 51–90 (2003). https://doi.org/10.2307/30036519.
43. Nick, B., Eliezer, Y.: The Ethics of Artifical Intelligence. Cambridge Handb. Artif. Intell. 47, (2011). https://doi.org/10.1016/j.mpmed.2018.12.009.
44. Moon, Y.: Intimate Exchanges: Using Computers to Elicit Self-Disclosure From Consumers. J. Consum. Res. 26, 323–339 (2000). https://doi.org/10.1086/209566.
45. Pavlou, P.A., Gefen, D.: Building effective online marketplaces with institution-based trust. Inf. Syst. Res. 15, (2004). https://doi.org/10.1287/isre.1040.0015.
46. Wambsganss, T., Rietsche, R.: Towards designing an adaptive argumentation learning tool. In: 40th International Conference on Information Systems, ICIS 2019. p. 1 (2020).
47. Gregor, S., Hevner, A.R.: Positioning and Presenting Design Science Research for Maximum Impact. (2013).
48. Gläser, J., Laudel, G.: Experteninterviews und qualitative Inhaltsanalyse : als Instrumente rekonstruierender Untersuchungen. VS Verlag für Sozialwiss (2010).
49. Cooper, H.M.: Organizing knowledge syntheses: A taxonomy of literature reviews. Knowl. Soc. 1, 104–126 (1988). https://doi.org/10.1007/BF03177550.
50. vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., Cleven, A.: Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. Commun. Assoc. Inf. Syst. 37, 205–224 (2015). https://doi.org/10.17705/1cais.03709.
51. Webster, J., Watson, R.T.: Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Q. 26, xiii–xxiii (2002). https://doi.org/10.1.1.104.6570.
52. Følstad, A., Brandtzaeg, P.B.: Users' experiences with chatbots: findings from a questionnaire study. Qual. User Exp. 5, (2020). https://doi.org/10.1007/s41233-020-00033-2.
53. van de Poel, I.: An Ethical Framework for Evaluating Experimental Technology. Sci. Eng.

Ethics. 22, 667–686 (2016). https://doi.org/10.1007/s11948-015-9724-3.
54. Weller, A.: Transparency: Motivations and Challenges. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 23–40 (2019). https://doi.org/10.1007/978-3-030-28954-6_2.
55. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. In: FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. pp. 279–288. Association for Computing Machinery, Inc (2019). https://doi.org/10.1145/3287560.3287574.
56. Spanoudakis, G.: Plausible and adaptive requirement traceability structures. In: ACM International Conference Proceeding Series. pp. 135–142 (2002). https://doi.org/10.1145/568760.568786.
57. Rothenberger, L., Fabian, B., Arunov, E.: Relevance of Ethical Guidelines for Artificial Intelligence - A Survey and Evaluation. Eur. Conf. Inf. Syst. ECIS 2019. 0–11 (2019).
58. Sharma, S., Henderson, J., Ghosh, J.: CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models. (2019).
59. Millar, J., Barron, B., Hori, K., Finlay, R., Kotsuki, K., Kerr, I.: Accountability in AI Promoting Greater Societal Trust. G7 Multistakeholder Conf. AI. 16 (2018).
60. Yan, M., Castro, P., Cheng, P., Ishakian, V.: Building a chatbot with serverless computing. In: Proceedings of the 1st International Workshop on Mashups of Things and APIs, MOTA 2016. Association for Computing Machinery, Inc (2016). https://doi.org/10.1145/3007203.3007217.
61. Dignum, V.: Responsible Artificial Intelligence: Designing AI for Human Values. ICT Discov. 1–8 (2017).
62. Cohn, M.: User Stories Applied For Agile Software Development. (2004).
63. Gregor, S., Chandra Kruse, L., Seidel, S.: Research perspectives: The anatomy of a design principle. J. Assoc. Inf. Syst. 21, 1622–1652 (2020). https://doi.org/10.17705/1jais.00649.
64. Wambsganss, T., Molyndris, N., Söllner, M.: Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach. In: 15th International Conference on Wirtschaftsinformatik. , Potsdam, Germany (2020). https://doi.org/10.30844/wi_2020_c9-wambsganss.
65. Wambsganss, T., Niklaus, C., Söllner, M., Handschuh, S., Leimeister, J.M.: A Corpus for Argumentative Writing Support in German. In: 28th International Conference on Computational Linguistics (Coling) (2020).
66. Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Leimeister, J.M., Handschuh, S.: AL : An Adaptive Learning Support System for Argumentation Skills. In: ACM CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2020).
67. Gregory, R.W., Muntermann, J.: Research Note: Heuristic Theorizing: Proactively Generating Design Theories, https://www.jstor.org/stable/24700315, (2014). https://doi.org/10.2307/24700315.