

International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

The Value-Sensitive Conversational Agent Co-Design Framework

Malak Sadek, Rafael A. Calvo & Céline Mougenot

To cite this article: Malak Sadek, Rafael A. Calvo & Céline Mougenot (2025) The Value-Sensitive Conversational Agent Co-Design Framework, International Journal of Human–Computer Interaction, 41:15, 9533-9564, DOI: 10.1080/10447318.2024.2426737

To link to this article: https://doi.org/10.1080/10447318.2024.2426737

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

Published online: 25 Nov 2024.

Submit your article to this journal

Article views: 2537

View related articles

View Crossmark data

Citing articles: 6 View citing articles

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=hihc20

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
2025, VOL. 41, NO. 15, 9533–9564
https://doi.org/10.1080/10447318.2024.2426737

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

The Value-Sensitive Conversational Agent Co-Design Framework

Malak Sadek, Rafael A. Calvo, and Céline Mougenot

Dyson School of Design Engineering, Imperial College London, South Kensington, London, UK

ABSTRACT

Conversational agents (CAs) are rapidly advancing across industry and academia and it is crucial to consider the values embedded within these systems. Value-sensitive design practices have benefited AI-based systems, but have not yet been widely applied to CAs. This paper introduces the Value-Sensitive Conversational Agent (VSCA) Framework. The framework uses collaborative design (co-design) activities to guide CA creators and CA users to collaboratively create three key artefacts that elicit CA users' values and are technically useful for CA creators to drive implementation forward, resulting in value embodied CA prototypes. The paper presents the practical framework and toolkit, followed by a mixed-method evaluation through design workshops, semi-structured interviews, and a comparative survey. Results show that the framework and toolkit increase CA creators' value-sensitivity, empower CA users, enhance collaboration, and produce value-embodied prototypes. Based on this work, we offer 14 guidelines to practically support value sensitivity in CAs.

KEYWORDS

AI; conversational agent; value-sensitive design; co-design

1. Introduction

Artificial Intelligence (AI) innovation shows no sign of slowing down, despite numerous warnings from experts. Conversational agents (CAs) are AI-powered systems delivering conversational experiences through voice or text (Deloitte Digital, 2019). They have seen exponential interest, research, and use (Wahde & Virgolin, 2022), recently driven by Large Language Models such as ChatGPT, earning them the description "Trojan horses for AI" (Russell, 2019, p. 68).

CAs exhibit biases present in other AI systems, e.g., voice technologies have shown biases against women (West et al., 2019), and users have reported having to sound "intentionally white" for recognition (Byrd, 2020). Subsequent harms and risks to users are difficult to identify, and even more difficult to measure (Jacobs & Wallach, 2021; Weidinger et al., 2023). AI-powered CAs are now used in critical domains such as healthcare, given their numerous benefits around cost reduction, workflow efficiency and patient communication (Xu et al., 2021; Siddique & Chow, 2021). Additionally, LLM-powered CAs are now unlocking even more use-cases within these domains than ever before (Chow et al. (2024) identify up to 18 use-cases in healthcare).

These biases, along with CAs' increasing use and influence, necessitates a deeper understanding and managing of values embedded within them. Practical mechanisms involving diverse stakeholders to collect values (Dantec et al., 2009) and embed them into technology (Manders-Huits, 2011; van de Poel, 2020) are needed, but currently limited in CA design practices (Byrd, 2020; Følstad et al., 2021; Showkat & Baumer, 2022; Sadek et al., 2023a).

Accordingly, our research re-imagines the current CA design process by (i) leveraging collaborative design (co-design) with diverse stakeholders (Sanders & Stappers, 2008) and (ii) focusing on value-sensitive outcomes (Friedman et al., 2006). Co-designing CAs aims to leverage cited benefits for co-designing technology in general (Sanders & Stappers, 2008; Hayes, 2020), and CAs specifically (Sadek et al., 2023a), as well as expert recommendations to do so (Chen et al., 2020; Lewandowski et al., 2021), and the necessity of involving stakeholders to elicit and verify embedded values. Creating value-sensitive CAs stems from the need to control value-based decision making in AI systems (Christian, 2020), the importance of value considerations in AI (Umbrello & Van de Poel, 2021), and CAs (Wambsganss et al., 2021), and the lack of emphasis on values in co-designing CAs (Sadek et al., 2023a). These two goals form the research basis of this work.

Our driving research question has been "how can we support and enable the co-design of value-sensitive conversational agents?". The aims of this work were therefore to:

1. Identify barriers and requirements for co-designing value-sensitive CAs.
2. Design and implement a practical framework to facilitate co-designing value-sensitive CAs.
3. Explore the framework's impact on CA prototypes produced using a mixed-method evaluation with teams building CAs.

CONTACT Malak Sadek m.sadek21@imperial.ac.uk Dyson School of Design Engineering, Imperial College London, South Kensington, SW7 2BX, London, UK

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

In this paper we present the Value-Sensitive Conversational Agent Co-Design Framework (VSCA) and its operationalisation through a toolkit (Figure 1). We then report two case studies of industry CA professionals using it in co-design workshops with stakeholders, semi-structured interviews exploring participants' experiences and perceptions, and a survey assessing the value-sensitivity of workshop outcomes.

In this context, values refer to notions and aspects that are important to people and that help them define how they want to live their life and distinguish "right" from "wrong" (Friedman & Hendry, 2019). Values are often confused with similar concepts such as ethics, morals, and principles. While the Stanford Encyclopedia of Philosophy offers a detailed explanation of their distinctions, the main differences in this context is that values are internalised on an individual level, as opposed to a community or societal level, are subjective rather than objective "truths", and can change over time. Values also differ from goals where values describe *how* someone wants to live their life (e.g., with "independence"), whereas goals outline *what* to do to achieve this state (e.g., learn about a bank's savings plans to make informed financial decisions for themselves). Several methods and taxonomies for categorising different types of values exist (Fuchsberger



Figure 1. The VSCA toolkit which operationalises the VSCA framework. Within the (a) enclosing box, the toolkit consists of (b) a guidebook with instructions, (c) a set of canvases to carry out activities, (d) information sheets to help create meaningful and relevant outputs, and (e) blank decks of cards used during the first activity.

et al., 2012; Schwartz et al., 2012; Yurrita et al., 2022). In relation to technology, the distinction is often made between "functional" values, such as privacy, safety, security, and others that can translate into functional requirements for a system; and "non-functional" values that are more social, epistemic or contextual in nature and do not necessarily directly translate to specific functions or features, such as empathy, curiosity, and dignity (Fuchsberger et al., 2012). An example of a functional value that came up during this work is "safety & security", which study participants defined as "taking the necessary steps to ensure that the user's data is safe and secure, while also making this clear to the user." An example of a non-functional value implicated in this work is "friendliness & playfulness", which participants defined as "feeling that the experience is friendly and making the conversation an enjoyable experience, regardless of the topic discussed or the outcome of the conversation."

2. Related work

This section explores some of the discourse and literature that has informed the creation of the VSCA framework. First, the need for a practical framework over other forms of interventions is discussed. Secondly, the use of boundary objects during AI systems design is explored, as the framework revolves around the creation of three boundary objects that embody stakeholders' values and can drive CA implementation forward. Thirdly, achieving value-sensitivity is discussed; this discussion helps contextualise the framework's rooting in value-sensitivity. Finally, the requirements for co-designing value-sensitive CAs are presented, as determined by previous work, in order to set the stage for the framework's objectives and structure.

2.1. Socio-technical interventions for AI systems

Abstract and technically-focused guidelines and recommendations for AI systems have faced several criticisms around: their abstract nature leaving interpretation up to practitioners (Whittlestone et al., 2019; Deng et al., 2023), the difficulty of measuring their impact (Hagendorff, 2020) and the possibility that they may have no impact at all (McNamara et al., 2018), their tendency to promote passive compliance over active consideration of socio-technical aspects (DigitalCatapult, 2020; Wachter & Mittelstadt, 2019), their over-formalisation of AI ethics (Ali et al., 2023), the over-saturation of principles for AI (Jobin et al., 2019) and CAs (Murad et al., 2023) making selection difficult, and their ineffectiveness at supporting value-related considerations (Palmer & Schwan, 2024).

Instead, researchers have called for concrete practices and toolkits that have socio-technical focuses (Harbers & Overdiek, 2022). Nevertheless, such interventions have not focused on supporting value-sensitive outcomes (Sadek et al., 2023a) and are not widely used by industry CA practitioners (Sadek et al., 2023b). Furthermore, they are limited in generalisability (Elshan et al., 2022) and design process coverage (Sadek et al., 2023a). Design processes and frameworks can therefore be framed as a middle ground, as shown in Figure 2. While design processes have also varied in their support for value-based considerations (Sadek, Calvo, et al., 2024), frameworks offer a middle ground. They benefit from broader perspectives and coverage than isolated tools (Morley et al., 2023), while concretely structuring activities and avoiding pitfalls of overly general guidelines. Frameworks can integrate into existing practitioner processes without disruptive redesigns, facing less adoption resistance (Morley et al., 2023). Experts call for such socio-technical frameworks for AI (Morley et al., 2023) and CAs (Lewandowski et al., 2021). We therefore selected a practical framework as an intervention in response to this discourse.

2.2. The use of boundary objects in designing AI systems

Boundary objects help build shared understanding across diverse participants (Pennington, 2010; Cabitza & Simone, 2012; Arrighi & Mougenot, 2019). In interdisciplinary settings, they serve as (i) tangible affordances of knowledge produced (Pennington, 2010; Cabitza & Simone, 2012; Arrighi & Mougenot, 2019), (ii) documentation and bases for future decision-making, and (iii) mediums for supporting discussions (Vines et al., 2013). Design workshops often involve creating or using boundary objects (Vines et al., 2013; Steen, 2013; Jonas, 2018). Tangible objects like canvases, cards, and sticky notes are ideal for quick ideation (Boujut & Blanco, 2003).

AI practitioners have mentioned needing boundary objects to aid collaboration, interdisciplinary communication, and establish common language (Lee, 2020; Yildirim et al., 2022; Kayacik et al., 2019; Martin, 2020). They find current boundary objects such as PowerPoint presentations time-consuming and lacking sufficient documentation (Piorkowski et al., 2021). The use of boundary objects has also been recommended (Lewandowski et al., 2021) and explored (Axelsson et al., 2022) in CA design. Recent work has suggested creating boundary objects for various AI stakeholders (Yildirim et al., 2023) and that data from non-technical stakeholder data can act as a boundary object and facilitate collaboration with technical stakeholders around AI-based systems (Yoo et al., 2024). Thus, the framework guides stakeholders in collaboratively designing boundary objects throughout the CA design process.

2.3. Achieving value-sensitivity

Values significantly influence users' perceptions of different technologies (Kim & Kim, 2024; Ashfaq et al., 2021) and the technologies they choose to use (Aguiar-Díaz & Ruiz-Mallorquí, 2024), particularly when the values embedded in a system align with those of users and stakeholders (Yokoi & Nakayachi, 2021). Value-Sensitive Design (VSD) has long advocated for the integration of stakeholders' values into technology development on conceptual, empirical, and technical levels (Friedman et al., 2006). This approach has been successfully applied to diverse technologies, including Virtual Reality (VR) (Sonntag et al., 2023) and mHealth



Figure 2. Frameworks can harness the benefits of guidelines & recommendations and design tools & activities while mitigating some of their drawbacks. This has especially been called for in terms of supporting socio-technical practices for AI (Morley et al., 2023) and CAs (Lewandowski et al., 2021).

applications (Almutairi et al., 2024). With the advance of LLMs, such as GPT, the need to align new CAs with their diverse stakeholders' values increases further (Chen et al., 2023). Nevertheless, recent work has found LLMs to reflect the values of a small segment of the global population (Benkler et al., 2023; Durmus et al., 2024; Atari et al., 2023), with each commercially available LLM harboring sets of distinctly different values (McGrath, 2024; Yi et al., 2023).

Considering values and embedding them within technologies is not exclusive to design disciplines. In fact, patients' values play a significant role in medical decision aids (Fagerlin et al., 2013) and domain experts' values have helped shape effective healthcare interventions (Strikwerda et al., 2022; Faber et al., 2024). Additionally, highlighting specific values such as "safety", "sustainability", and "reliability" has been found to impact the effectiveness of strategies for improving environmental practices (Tanner & Feltz, 2022). Moreover, stakeholder values have played a key role in improving outcomes across a vast number of domains and fields, from manufacturing (Vernim et al., 2022) to law and policy-making (Umbrello, 2021). Examining how values are elicited and engaged with across these diverse fields can also yield valuable insights into how to verify and measure value-sensitivity. For example, technology adoption frameworks, such as the Unified Theory of Acceptance and Use of Technology (UTAUT), often assess a number of implicated values, such as perceived "usefulness" (Venkatesh, 2022). Similarly, a standardised framework or survey could help assess different values, and especially non-functional ones, which is a research avenue worth exploring in the future.

In the context of AI-based systems, recent research has focused on applying VSD to these systems (Umbrello & Van de Poel, 2021), exploring public sentiment on values related to AI ethics (Ikkatai et al., 2022), and on considering whose values should be embedded in AI systems (Christian, 2020) and how this embedding might be achieved (van de Poel, 2020). These works are rooted in the premise that a focus on values can bring numerous benefits to the development of AI systems, such as promoting reflexivity among practitioners (Umbrello & Van de Poel, 2021), enhancing credibility and accountability (Stray et al., 2020; Park et al., 2022), and fostering a society-in-the-loop approach with increased community participation (Rahwan, 2018). Despite advancements in evaluating the usability and user experience of Conversational Agents (CAs) (Bubaš et al., 2023), developing design guidelines for value-sensitive CAs (Wambsganss et al., 2021), and creating value-sensitive CAs with specific stakeholder groups (Showkat & Baumer, 2022), there has been limited work on adapting existing and effective value-centric frameworks, such as VSD, specifically to CAs. As such, this work leverages existing literature on adapting VSD for AI-based systems and develops a framework for bringing VSD principles to CA design.

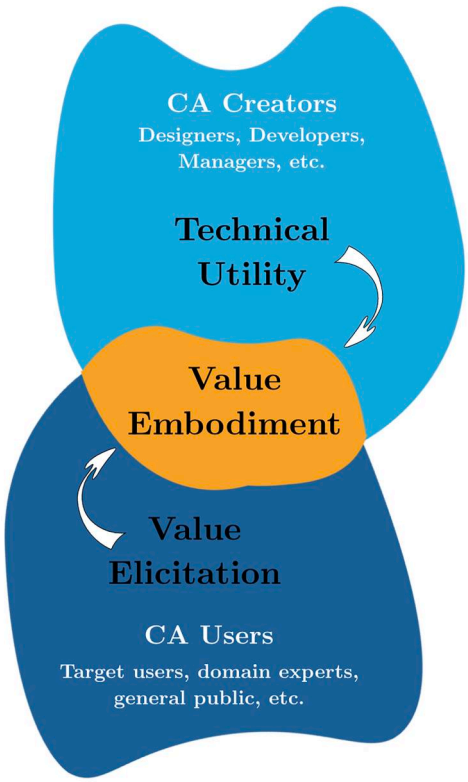


Figure 3. A model of the requirements needed to reach value embodiment which is framed as a collaborative effort between CA teams and external stakeholders. The values of external stakeholders need to be elicited, and collaboration outcomes need to have a technical utility in order for the CA team to contribute towards value embodiment.

2.4. Requirements for co-designing value-sensitive conversational agents

Based on prior research conducted[1], a model for co-designing value-sensitive CAs can now be conceptualised, as shown in Figure 3.

2.4.1. Reaching value embodiment

VSD offers little guidance regarding how to ensure that chosen values have been embedded into a technology (Manders-Huits, 2011). Alternatively, van de Poel (2020) defines three types of values present during the creation of an AI system:

- **Intended Values:** These are the values present during the beginning of the project that the creators intend to embed in the system.

- **Embodied Values:** These are the values present in the intermediate artefacts and prototypes created during the design of an AI system.
- **Realised Values:** These are the values that emerge after the system is deployed and as a result of its usage (whether it is used as intended or not).

They recommend focusing on "embodied values", as opposed to intended values that manifest before the design process begins (which may not propagate along the process) or realised values that emerge after it has been built and is in use (which are dependent on outside contextual factors beyond the scope of the creators). In order to achieve meaningful value embodiment, both "CA users" (including target users, the general public, etc.) and the "CA Creators" responsible for the CA (including designers, developers, managers, domain experts, etc.) must collaborate. CA users' values must be collected and understood (**value elicitation**), and for CA creators to meaningfully engage with these values, the outcomes of the collaboration need to have **technical utility** - i.e., they need to be understandable and useful from a technical perspective in order to be accepted by practitioners and have the ability to drive CA implementation.

2.4.2. Value elicitation

Prior work has distinguished between using a top-down approach where pre-defined values from an existing source are used, and a bottom-up approach where contextual values are elicited from the stakeholders of a technology (Vera & Muller, 2019). A bottom-up approach is generally favored as it avoids several limitations of using pre-defined values, such as their lack of generalisability across groups of people and contexts (Jakesch et al., 2022) and their abstract nature, which makes them difficult to practically implement (Pommeranz et al., 2012). The use of a bottom-up approaches towards value elicitation has also been found to be effective in the context of AI systems (Verdiesen & Dignum, 2022), where creators' values can differ from those using the systems and affected by them (Jakesch et al., 2022).

2.4.3. Technical utility

For many co-design activities, it is unclear to technical practitioners how to use the outcomes produced. There is usually an extra "translational work" (Wong et al., 2022, p. 17) needed in order for outcomes to become "actionable insights" (Deng et al., 2023, p. 11) that are useful from a technical perspective, in terms of practitioners being able to incorporate them into their workflows (Wong et al., 2022). This contributes to technical practitioners having negative attitudes towards socio-technical design aspects (Manders-Huits & Zimmer, 2009; Dindler et al., 2022) as they view these considerations as burdens or extra work that they must take on, in addition to their usual technical responsibilities (Morley et al., 2023).

In summary, literature has shown that a practical, socio-technical framework can offer operationalisability, coverage, and generalisability to support value-related considerations during CA design; that co-designing boundary objects is needed for facilitating interdisciplinary collaborations around AI-based systems; that CA design can benefit from the application of VSD principles; and that value embodiment can be achieved through eliciting CA users' values and providing CA creators with outcomes that have technical utility. These insights inform our research question: "How can we support and enable the co-design of value-sensitive conversational agents?" Our framework focuses on co-designing boundary objects at different CA design stages to elicit CA users' values and provide technical utility to CA creators, enabling the co-design of value-embodied prototypes. Following the framework and toolkit's creation (Section 3), they are evaluated through design workshops with industry CA practitioners (Section 4), a comparative survey (Section 5), and semi-structured interviews with workshop participants (Section 6).

3. The VSCA framework

The structure of the framework, its operationalisation as a toolkit, and a breakdown of its activities are presented in this section.

3.1. Framework structure

The main element of the framework is facilitating the co-design of three boundary objects by CA creators and CA users through three activities. The key ingredient in these boundary objects is the concurrent focus on both highlighting elicited values and providing technical utility. When combined, these provide a value embodied CA prototype. This is a stark difference to how the majority of CAs are currently designed: in isolation (Elshan et al., 2022; Sadek et al., 2023b), and with no focus on stakeholders' values (Sadek et al., 2023a). These activities and the resulting boundary objects are shown in Figure 4. In the figure, they are mapped to relevant VSD investigations (Friedman et al., 2006), proposed steps for embedding values in technology (Flanagan et al., 2008), and a conceptual framework for applying VSD to AI systems (Umbrello, 2018) from all of which they were derived and adapted to CA design. Each activity includes an element for value elicitation, technical utility, and a value-lever discussion task from previous works and established literature, which is referenced in the figure.

The design activities which lead to the co-design of the three mentioned boundary objects act as value levers. Value levers are activities which facilitate discussions and actions around values during a project (Shilton, 2013). These can be any activity, including gaining funding or simply working in an interdisciplinary team (ibid). Previous work has shown that design activities can function as value levers (Ahn et al., 2021) and that the use of value levers is effective in the context of AI systems (Varanasi & Goyal, 2023; Winecoff & Watkins, 2022).

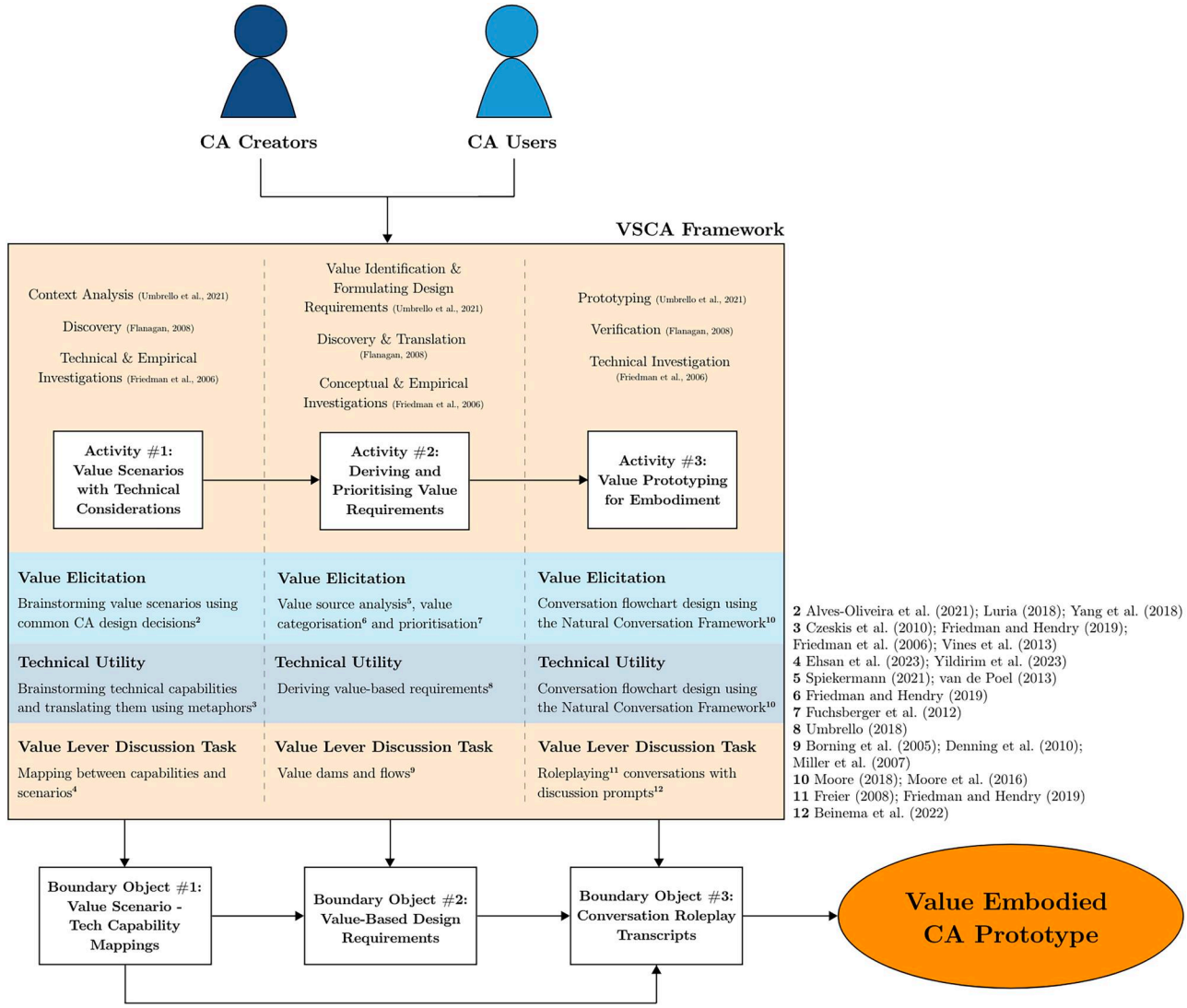


Figure 4. A diagram showing the activities and boundary objects included in the VSCA framework.

As shown in Figure 4, recommendations and strategies from existing literature have shaped each stage and activity in the framework. Additionally, the several aspects of the framework and toolkit are influenced by a number of related work around VSD and designing CAs. These influences are listed in Table 1. Particularly, it is worth noting that related approaches have looked at working with domain-specific values (Dickhaut et al., 2024; Faber et al., 2024), as well as culture-specific values (Reinecke & Bernstein, 2013; Janson et al., 2022). This framework aims to learn from these approaches and extend them to the consideration of *any* contextual values, which are likely to depend on the stakeholders involved in using the toolkit. For example, the inclusion of domain experts might lead to the elicitation of more domain-specific values, while the inclusion of general public users for a commercial CA might elicit more diverse values spanning across several cultures. In such cases and contexts, the use of this framework as a guide to effectively eliciting values and embedding them in CAs in verifiable ways can take place alongside the aforementioned approaches that provide more targeted support. In particular, when used modularly, the activities and tools provided in the toolkit can help carry out and provide documentation for several overlapping steps across the mentioned approaches around identifying values, deriving value-based design requirements, and tailoring design features around implicated values (Dickhaut et al., 2024; Janson et al., 2022; Reinecke & Bernstein, 2013).

Combining these approaches and tools allows CA creators to reap the benefits of each one, while also helping to consolidate the current fragmentation in CA research and development (Elshan et al., 2022; Wang et al., 2023).

Nevertheless, the VSCA Framework and Toolkit also differs from previous work in several key ways that are recommended across the literature, but have not been systematically and concurrently incorporated into practical frameworks and approaches, nor focused on value alignment, as of yet:

- Empirically verifying that outcomes are value-sensitive by leveraging the concept of translating intended values into embodied values (van de Poel, 2020; Manders-Huits, 2011; Flanagan et al., 2008);

Table 1. Framework aspects influenced by key literature around frameworks and approaches for VSD and CA design.

Aspect	Source
Producing documents detailing outcomes for transparency and accountability.	(Dickhaut et al., 2024; IEEE 7000, 7000, 2021).
Considering consequences and risks for outcomes produced.	(Dickhaut et al., 2024; IEEE 7000, 7000, 2021).
Deriving design requirements from important values and use-cases.	(Dickhaut et al., 2024; Janson et al., 2022).
Concurrently focusing on both technical and non-technical aspects.	(Dickhaut et al., 2024; Wolff et al., 2022).
Considering key design decisions within a CA beyond conversational content.	(Elshan et al., 2023; Wolff et al., 2022).
Mapping technical aspects to user and business aspects.	(Elshan et al., 2023).
Starting with value identification before deriving requirements.	(Elshan et al., 2023; IEEE 7000, 7000, 2021) (Janson et al., 2022).
Translating individual values into design features and requirements.	(Janson et al., 2022; Reinecke & Bernstein, 2013).
Providing coverage for several major design phases to avoid ambiguity.	(Sadek, Calvo, et al., 2024).
Providing tangible tools for clarity and practicality.	(Dickhaut et al., 2024; IEEE 7000, 7000, 2021).
Systematically involving stakeholders through the use of engaging and accessible activities.	(Sadek et al., 2023a; Sadek, Calvo, et al., 2024).
Including prototype evaluation criteria around value alignment.	(Umbrello & Van de Poel, 2021; Zhu et al., 2018).
Providing prompts, cues, and examples for guidance.	(Sadek, Constantinides, et al., 2024).
Exploring values deeply through value-based design fictions.	(Vera & Muller, 2019).

- Advocating for, and facilitating, diverse and interdisciplinary stakeholder participation (Lewandowski et al., 2021);
- Focusing on the specificities of an applied AI-based technology, as opposed to generalising across all AI (Elshan et al., 2022);
- Addressing practitioner reluctance to use such interventions by focusing on technical utility (Manders-Huits & Zimmer, 2009; Morley et al., 2023);
- Offering practical guidance and support through a dedicated toolkit (Harbers & Overdiek, 2022; Dantec et al., 2009);
- Producing outcomes in the form of boundary objects that:
  - Facilitate interdisciplinary collaboration by producing shared knowledge (Pennington, 2010; Arrighi & Mougenot, 2019; Yildirim et al., 2022),
  - Provide tangible outcomes to center co-design efforts around (Vines et al., 2013; Steen, 2013; Lee, 2020; Yildirim et al., 2022),
  - Increase accountability by providing documentation (Sleeswijk, 2018; Koesten et al., 2019; Zhang et al., 2020; Custis, 2021),
  - Propagate stakeholders' values forward to later stages of the design process (Monteiro & Salgado, 2023; Yildirim et al., 2023; Dindler et al., 2022).

3.2. Framework operationalisation

The framework is operationalised using a design toolkit (shown at the beginning of the paper in Figure 1, and broken down in the following section). The toolkit provides templates for each of the boundary objects that facilitates their creation and standardises their form. The goal is to provide a concrete and practical means for following the framework that CA creators can use in a design workshop setting to facilitate team activities and actualise the needed boundary objects. The toolkit is composed of a guidebook, a set of canvases for the team activities and the boundary objects, a set of blank cards to be filled out for the first boundary object, and information sheets that provide helpful information during different activities in order to keep participants' ideas focused and relevant.

3.3. Framework activities

Each of the three activities has an identical structure which is based on the Value-Sensitive Action Reflection model (Yoo et al., 2013). During each activity, participants are split into CA creators and CA users. Each team works on a task separately, then both teams come together to combine and consolidate their work into the given boundary object, while engaging in discussion and consensus-building. The tasks that each team works on for each activity and the underlying literature that underpins each task are shown in Figure 4.

3.3.1. Activity #1 - Value scenarios with technical considerations

Activity 1 involves a contextual analysis aimed at aligning stakeholders' values with available technological capabilities and opportunities for CAs. CA creators engage in brainstorming sessions to identify relevant technological capabilities, utilising metaphors to improve CA users' comprehension, drawing from research on anthropomorphism (Luria, 2018) and idea generation (Alves-Oliveira et al., 2021). These capabilities can include the ability to collect data through sensors or questions, the ability to infer different types of data from each other, the ability to do things on users' behalf using APIs, and so on. Simultaneously, CA users create value scenarios, which serve as a means of value elicitation and technological framing (Friedman et al., 2006). Additionally, stakeholders receive guidance on common CA design decisions to ensure their scenario are realistic and technically feasible. This indirect approach to value elicitation through the use of envisioning and other design activities (Lim & Kim, 2022) aligns with recommendations around surfacing and understanding of implicit values (Cockton, 2005, 2006).

Following individual tasks, each team will have written their ideas on a set of blank cards. Both CA creators and CA users will then come together and match capabilities cards with value scenario cards, facilitating discussion on unaddressed scenarios and underutilized capabilities. This process aims to ensure that technological decisions are driven by stakeholder needs and values, avoiding the pitfalls of technology-centric solutions that "[fit] people to technologies" (Vines et al., 2013) or focus on a "solution in search of a need" (Calvo & Peters, 2017). This consideration of human-side and technology-side aspects and bridging between them also has parallels in Yildirim et al.

(2023)'s findings of how AI practitioners work to "match AI capabilities with human needs", and the approach taken by the Conversational Design Institute (CDI)'s in their Conversation Design Canvas[2]. The key differences are a distinct focus on values, the facilitation of communication through the use of metaphors, and scenario scoping through common CA design decisions. Figure 5 shows the canvas and cards used.

3.3.2. Activity #2 - Deriving and prioritising value requirements

The objective of Activity 2 is to extract and categorize values from the "value scenario-technology capability" pairs created during Activity 1 and convert them into value-based design requirements for the envisioned CA. For the initial task conducted in two individual groups (CA creators vs. CA users), the CA users will engage in value categorisation based on the value scenarios produced during the previous workshop (Friedman & Hendry, 2019, p. 62). This categorisation is recommended in VSD as a method for more clearly identifying and resolving tensions (Friedman & Hendry, 2019, p. 68) and is done based on the Values in Action (ViA) approach (Fuchsberger et al., 2012) which builds upon the Theory of Consumption Values (TCV) (Hedman & Gimpel, 2010). The approach allows for taxonomising a wide range of values and understanding which value classes are of most relevant in different contexts, in order to prioritise those. Meanwhile, the CA team will write out value-based design requirements based on the use-case targeted, the technological capability utilised, and the underlying values involved. A number of previous works have derived design requirements for AI systems based on sets of ethical principles (Kroes & Van de Poel, 2020; van de Poel, 2013; IEEE 7000, 7000, 2021).

Both groups will then come together and share their outcomes. The group can then derive extra design requirements based on the categorised values and modify existing design requirements to explicitly mention the underlying values to be preserved in each case. Value Dams and Flows are then used, which have proven useful in other value-oriented studies for value-based analysis and selection of criteria/features/etc. (Borning et al., 2005; Denning et al., 2010; Miller et al., 2007; Czeskis et al., 2010). Figure 6 shows the canvas used by CA users during this activity.

3.3.3. Activity #3 - Value prototyping for embodiment

The objective of Activity 3 is to create a conversational prototype based on the value scenarios and value-based design requirements produced in the previous two activities. For the initial task conducted in two individual groups (CA creators vs. CA users), both teams will separately create conversation dialogue flows for any number of value scenarios created in Activity 1. In order to aid them in doing so, they will be provided with conversation patterns from the Natural Conversation Framework (Moore et al., 2016; Moore, 2018) to ensure participants' conversational

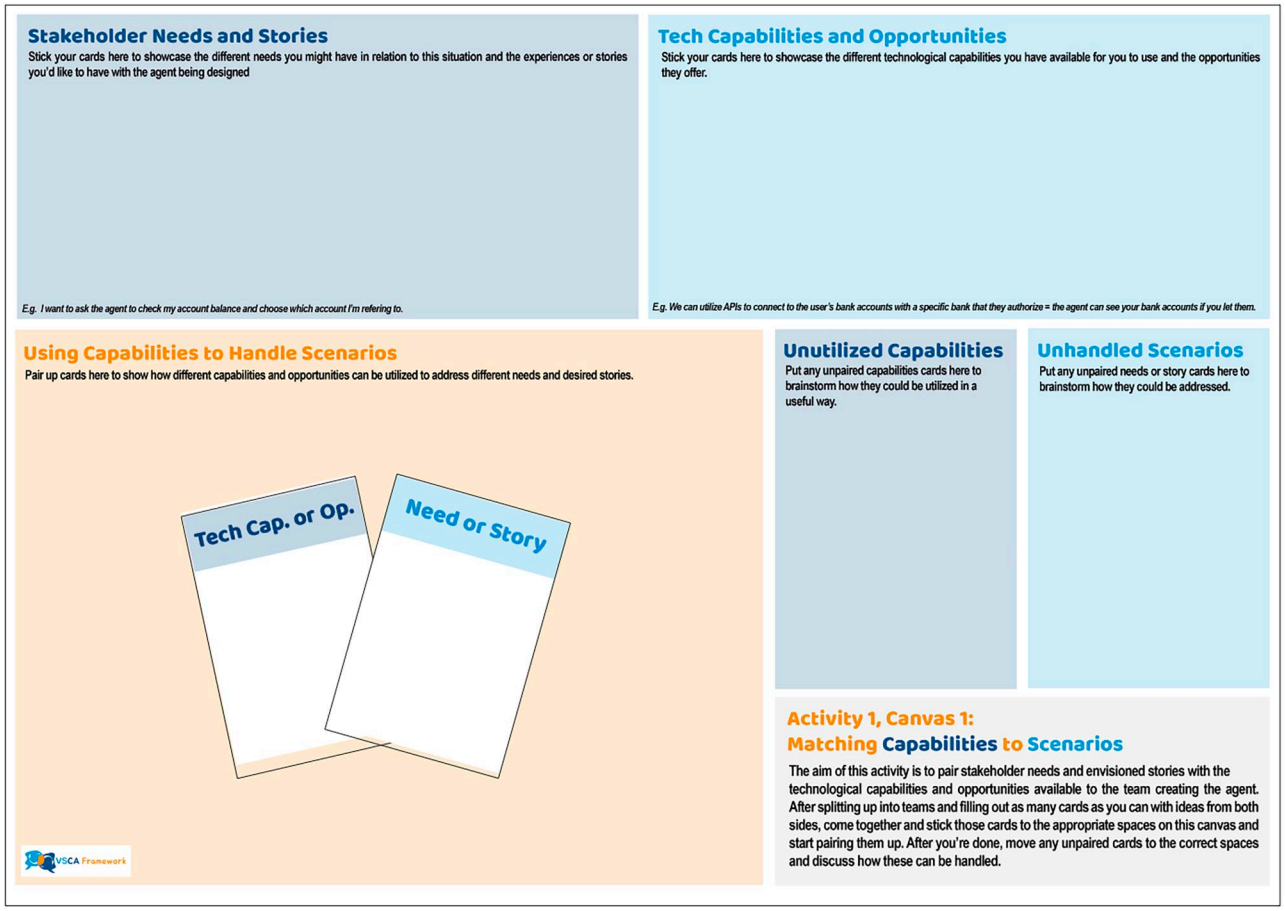


Figure 5. Activity 1 toolkit components: blank cards for brainstorming technological capabilities and value scenarios and a canvas with allocated spaces for capabilities, scenarios, matched pairs, unutilised capabilities, and unhandled scenarios.

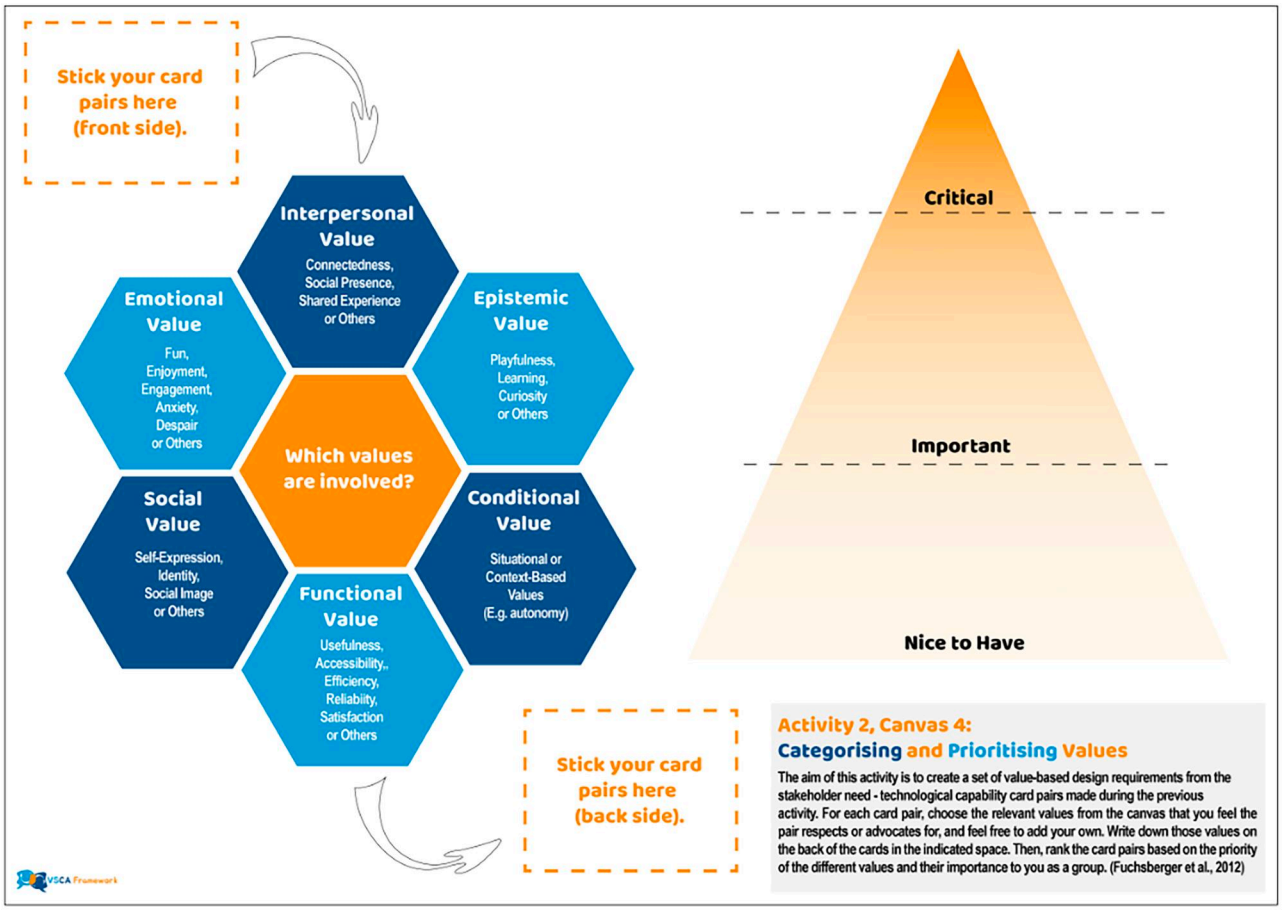


Figure 6. The canvas used by CA users during activity 2 to conduct value source analysis, categorisation and prioritisation using the backs of the cards used during activity 1 which have spaces for writing values and priority levels.

flows adhere to standards and are created in a form that is directly usable during implementation.

As the two groups come together, they will then roleplay CA creators' and CA users' different versions of a scenario, which is a very commonly used technique in CA co-design (Garg & Sengupta, 2020; Easton et al., 2019; Candello et al., 2020; Lee et al., 2017; Weber et al., 2021; Luria et al., 2020; Lopatovska et al., 2022). They can then discuss differences in content, tone of voice, agent personality, and so on using a set of discussion prompts provided. The transcripts of these roleplays form a prototype which embodies the learnings, elicited values, and outcomes of the workshops in a form usable by technical practitioners. These value-oriented prototypes (Friedman & Hendry, 2019, p. 62) contextualise the CA in real-world settings, and has been used in value-oriented investigations (Yoo et al., 2013) and personified agents (Freier, 2008). The resulting value-embodied boundary object takes the form of the roleplay transcripts, the notes of resulting discussions and changes, and the conversational flows. Figure 7 shows the toolkit components used during this activity.

4. Case studies - Co-design workshops

The flow of the evaluation protocol is shown in Figure 8 and Table 2. Accordingly, this section discusses the methodology of the design workshops themselves and the data collection and analysis in relation to the first aim. The second and third aims are explored in the following sections (Section 6 and Section 5 respectively).

In order to evaluate the VSCA framework and toolkit, two conversational agent co-design workshops were conducted with a variety of industry professionals. The workshops had three aims:

1. To observe and analyse CA creators' use of the framework and toolkit while co-designing CA prototypes with CA users (workshop recording and observations).
2. To produce two CA prototypes whose value embodiment can be evaluated and compared to two control prototypes (survey). As the framework recommends iteration over the prototyping stage, one CA prototype goes through one iteration (one workshop) and the other prototype goes through two iterations (two workshops).
3. To allow participants to experience the framework and toolkit as intended and then understand their perceptions and experiences (semi-structured interviews).

4.1. Methodology

The two workshops which used the VSCA toolkit took place on two separate dates three weeks apart in late 2023. These were followed by the control workshop which used the co-design activities featured in the framework (brainstorming,

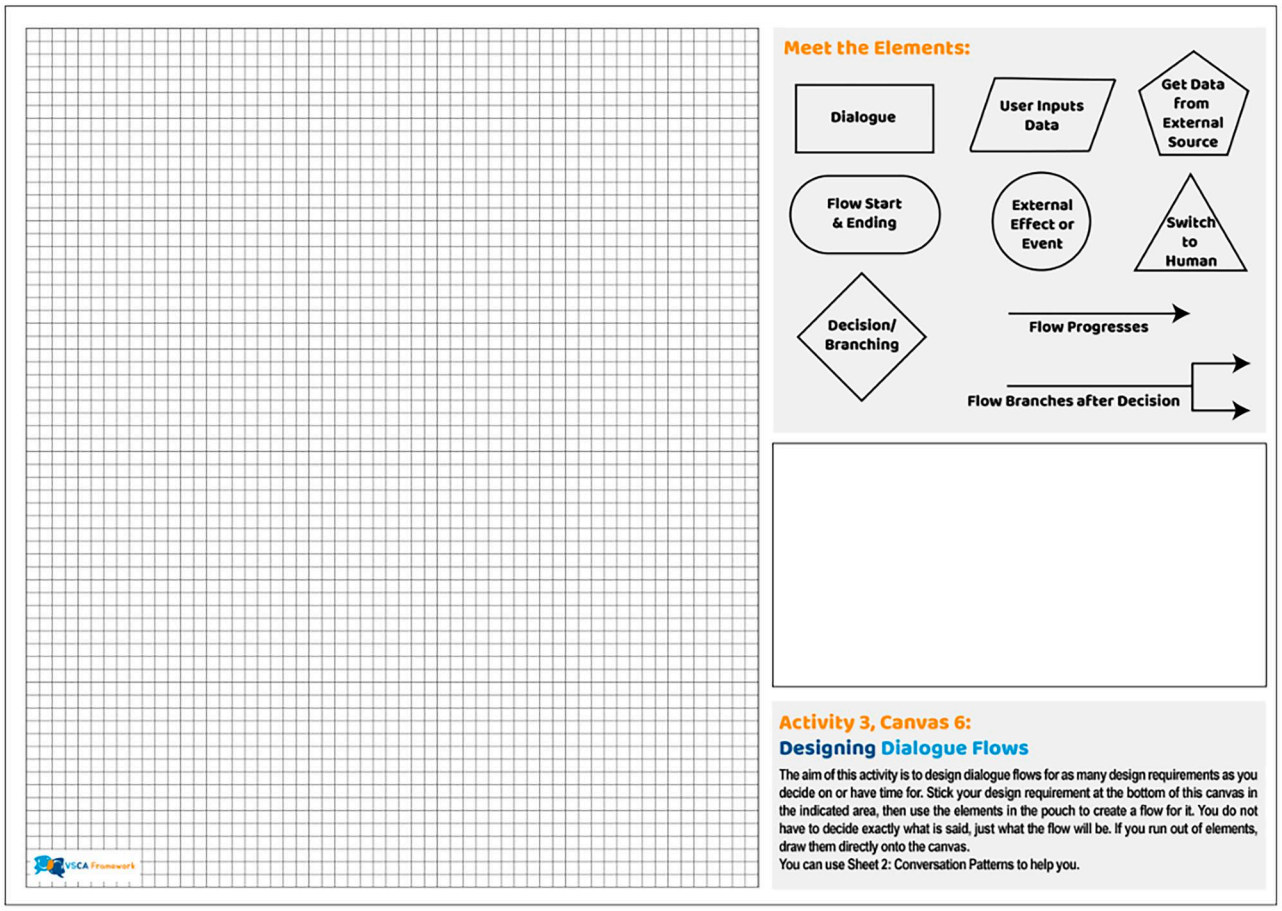


Figure 7. A canvas for drawing conversational flowcharts used by CA users and CA creators during activity 3.

co-writing design requirements, and roleplaying) but without the VSCA toolkit or any mention of the framework. All workshops were held online using Microsoft Teams and a Miro board where the toolkit was shared. Each workshop lasted three hours and were recorded and transcribed using Microsoft Teams. The first workshop was attended by 12 participants and the second workshop included 10 participants.

4.1.1. Workshop structure and activities

The three workshops had identical structures but different activities. Each workshop lasted three hours, divided into three 50-minute exercises with two 10-minute breaks. The VSCA workshops followed the activities and toolkit components described in Section 3, while the control workshop used brainstorming, co-writing design requirements, and roleplaying activities. Participants in the control workshop brainstormed freely using the same design brief as the VSCA workshops and then focused on two specific use-cases (budgeting and saving for a goal) for activities 2 and 3 to ensure comparability. For the control workshop, collaboration was promoted in the same way as the workshops with the VSCA Framework and Toolkit, where participants were told to work together in groups and were divided into CA creators and CA users at the start of each activity and then came together during the second half of each activity. The design brief, created in collaboration with an online community of conversation designers not participating in the study, contained all the necessary details for designing and prototyping a conversational agent.

A banking application was chosen as (i) the most common use-case for conversational agents is for customer service-related applications (RASA, 2021), (ii) it is an application that most participants are likely to be familiar with or have experience with, and (iii) it is an application where considering users' values is important but is perhaps less obvious or critical than other domains, such as healthcare, thus providing a middle-ground in terms of value prominence, in order to improve the generalisability of findings across domains.

4.1.2. Participants

Participants were recruited based on their professional experience through Linkedin. Participants were selected if they fit within one of two roles:

1. **Role 1 - CA Creators:** People with professional experience building or designing conversational agents This can include designing conversations, coding, and technical development.
2. **Role 2 - CA Users:** People who do not understand how conversational agents work or are built, but have a bank account or experience with banking services.

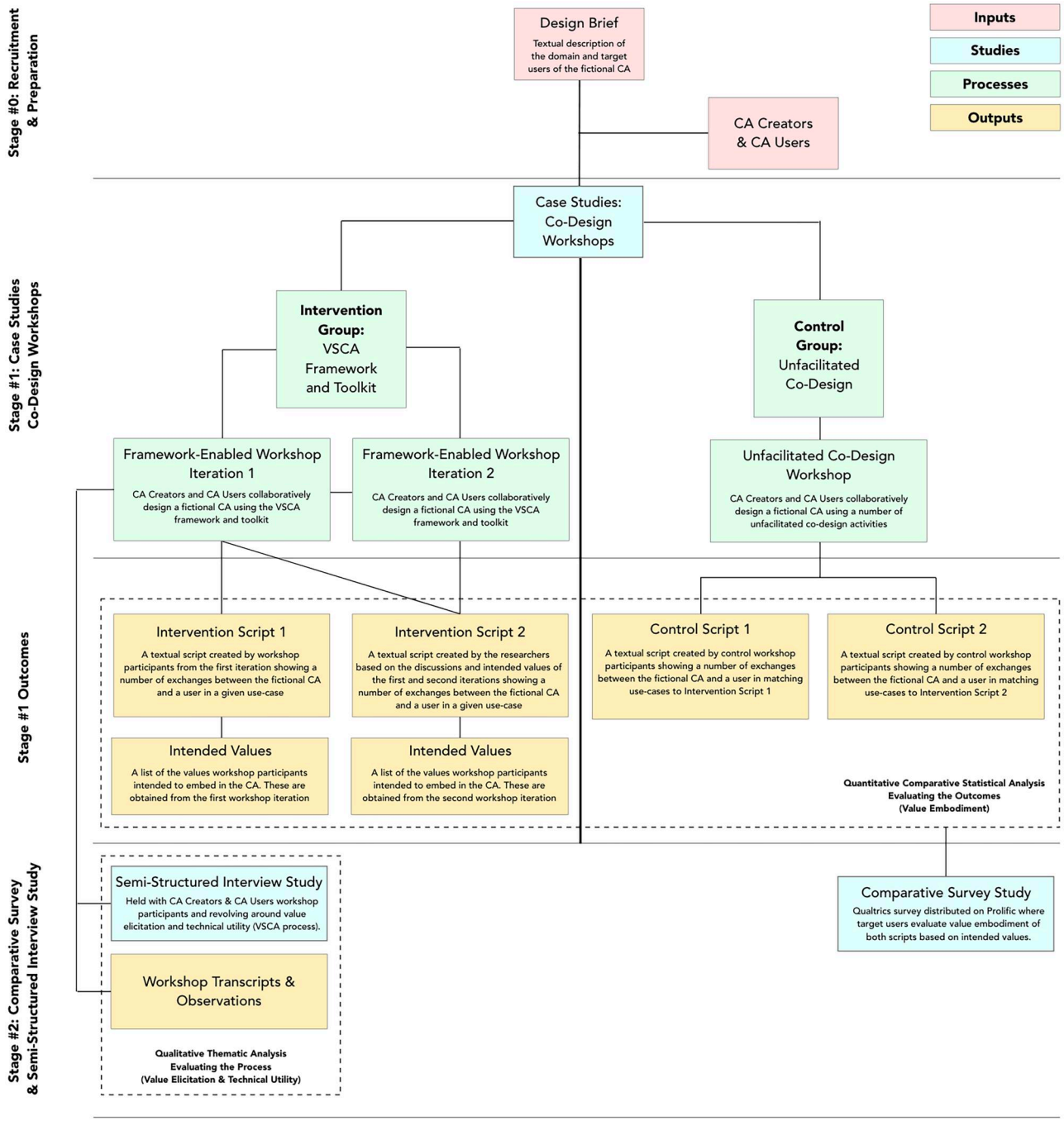


Figure 8. A flowchart outlining the three stages of the evaluation protocol for the VSCA framework and toolkit.

The first VSCA workshop included 7 participants in role 1 and 5 in role 2, with 6 females and 6 males, aged 22-40 (M = 29, SD = 7.3). The second VSCA workshop had 5 participants in role 1 and 4 in role 2, with 6 females and 4 males, aged 22-40 (M = 31, SD = 9.2). The control workshop included 4 participants in role 1 and 3 in role 2, with 5 females and 2 males, aged 22-40 (M = 29, SD = 5.6). Table 3 summarizes participants' demographics. Group sizes align with recommendations for effective engagement and brainstorming (Visser et al., 2005). Role 1 participation was voluntary, while role 2 participants received Amazon gift cards. Information sheets and consent forms were provided according to university ethics committee approval (reference 21IC7361).

To ensure comparability, participants across workshops had similar mean ages, gender compositions, and areas of expertise, experience in interdisciplinary collaboration, and no prior experience with frameworks related to human values.

4.1.3. Data collection & analysis

Data was collected during the workshops through recordings, transcriptions, and the outcomes produced by participants on Miro boards. In Activity 1, participants created sticky notes with potential use-cases and technological capabilities. Activity 2 generated sticky notes with values and design requirements. Activity 3 produced conversation flowcharts and scripts,

Table 2. A table summarising all the data collection and analysis activities conducted as part of this study.

Data name	Data type	Data source	Data objective	Metric
Workshop Recordings	Qualitative - Transcripts Thematic Coding	Co-Design Workshops	Analysing discussions and comments during the workshops	Process - Value Elicitation & Technical Utility
Workshop Participants' Interviews		Semi-Structured Interviews	Understanding participants' experiences and perceptions of the process & toolkit	
Survey Question 1 Responses	Quantitative - Rating Statistical Significance Testing + Qualitative - Value Mapping Analysis	Online Survey	The proportion that intended values represent out of all perceived embodied values	Outcome - Value Embodiment
Survey Question 2 Responses	Quantitative - Rating Statistical Significance Testing		Value alignment between intended and embodied values	
Survey Question 3 Responses			The extent to which intended values are aligned with target users' values	

Table 3. Workshop participants' demographics.

Participant ID	Age	Gender	Job title	Role
VSCA Workshop 1				
P111	24	Female	Start-Up Founder	Role 1
P112	40	Male	Conversational AI Consultant	Role 1
P113	25	Male	Machine Learning Engineer	Role 1
P114	40	Male	Conversational AI Consultant	Role 1
P115	25	Male	AI Ethicist	Role 1
P116	38	Female	Conversational Coaching Designer	Role 1
P117	37	Male	Creative Strategist	Role 1
P121	25	Male	Researcher	Role 2
P122	22	Female	Researcher	Role 2
P123	24	Female	Start-Up Founder	Role 2
P124	25	Female	PhD Student	Role 2
P125	23	Female	PhD Student	Role 2
VSCA Workshop 2				
P211	32	Female	Conversational AI Engineer	Role 1
P212	38	Male	Voice User Experience Designer	Role 1
P213	25	Female	Conversation Designer	Role 1
P214	49	Female	UX Design Manager	Role 1
P215	40	Female	Conversational AI Consultant	Role 1
P221	23	Male	Student	Role 2
P222	22	Female	Banking Design Engineer	Role 2
P223	29	Female	Researcher	Role 2
P224	29	Male	Researcher	Role 2
P225	20	Male	PhD Student	Role 2
Control Workshop				
P311	26	Male	Conversation Designer	Role 1
P312	27	Female	Conversation Designer	Role 1
P313	41	Male	AI Ethicist	Role 1
P314	28	Female	Conversational AI Developer	Role 1
P321	28	Female	Researcher	Role 2
P322	31	Female	Researcher	Role 2
P323	24	Female	Designer	Role 2

representing conversational agent prototypes. Figure 9 shows an example of these flowcharts with sticky note annotations indicating the intended values at different points.

The outcome of the VSCA workshops was two scripts between a CA and a user. One script was created by the participants of Workshop 1 (one iteration of the VSCA framework). The second script was started by Workshop 1 participants and updated by researchers based on Workshop 2 discussions and values, representing two iterations of the framework. This research design was chosen as the framework recommends iteration over the prototyping phase and so the performance of a script undergoing one iteration and another undergoing two iterations are compared. The control workshop also produced two scripts for use-cases matching those of the VSCA workshops (help with budgeting and saving for a goal).

To ensure practicality and applicability to real-world settings, and in line with recommendations to enhance the external validity of Responsible AI tools (Berman et al., 2024), the second VSCA workshop involved new participants. This simulated team changes and new stakeholder involvement, testing the framework under more challenging conditions and increasing the richness of the qualitative data collected. It also follows the recommendation of co-design in general to involve as many voices as possible (Sanders & Stappers, 2008).

The workshop transcripts and outcomes produced were analysed using thematic analysis alongside post-workshop interview transcripts. Accordingly, the results of both are presented and discussed in Section 5.

5. Semi-structured interview study - Exploring workshop participants' perceptions of value elicitation and technical utility (framework process)

The interview study qualitatively explored Role 1 (CA creators) and Role 2 (CA users) participants' experiences per Figure 3 in terms of the value elicitation that took place throughout the process and the technical utility of the outcomes produced. Given the focus on participants' experiences with the framework and toolkit, control workshop participants were not interviewed.

5.1. Methodology

Semi-structured interviews took place online within a week after each workshop, lasting 30 minutes. Participation was voluntary, with 8/12 from the first and 5/10 from the second workshop interviewed, totaling 13 (Table 4), in line with saturation recommendations (around 9-17 interviews) (Hennink & Kaiser, 2022). Their demographics are summarised in Table 4 using the same participant IDs as Table 3.

Interview questions differed by role (Appendix A). Given prior introductions and brevity, interviews started directly, with the Miro board link provided to participants to refresh their memory

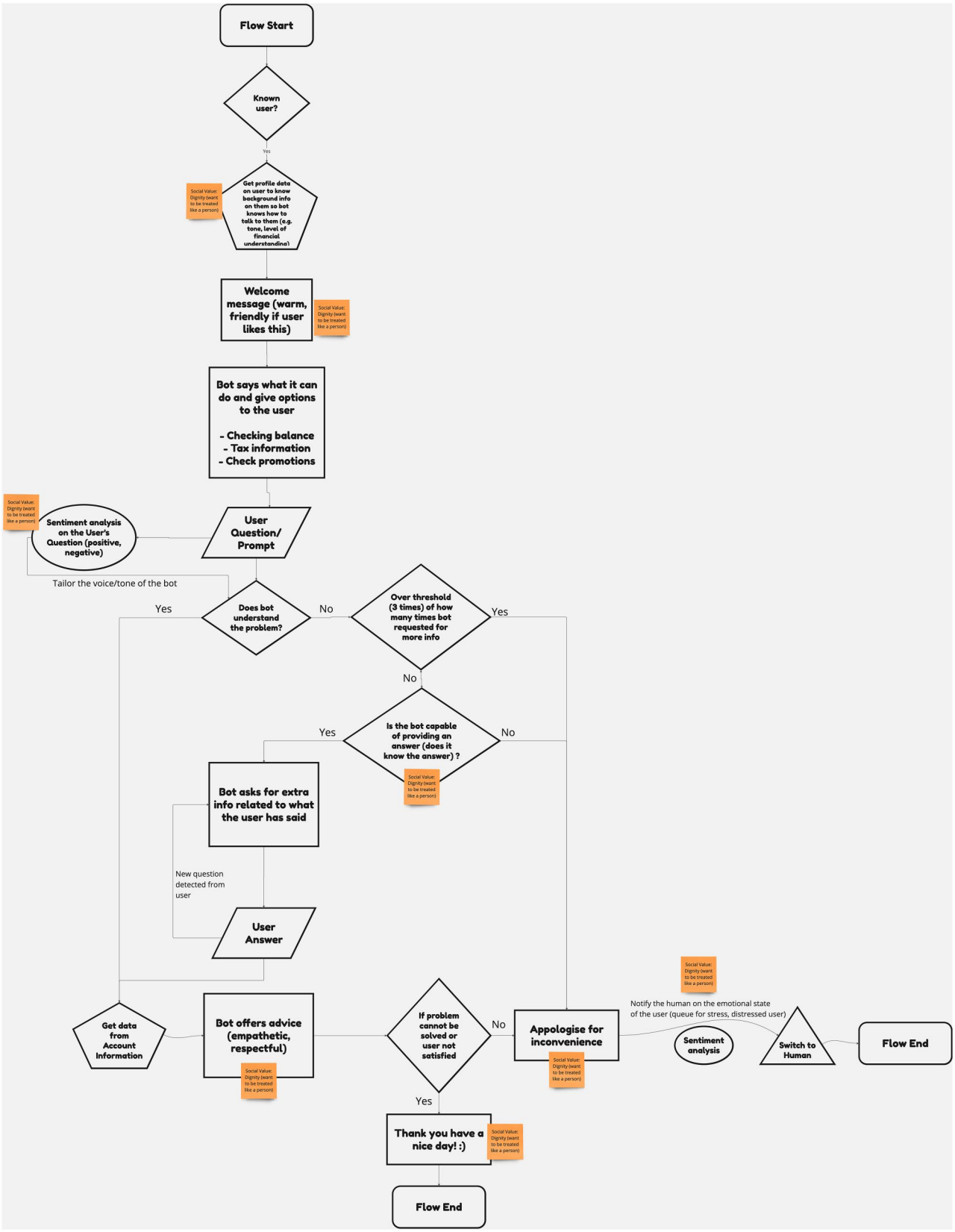


Figure 9. A flowchart of a sample conversation created by workshop participants during activity 3. The sticky notes annotate different intended values that the participants incorporated during the corresponding points of the conversation in their scripts.

Table 4. Interview participants' demographics.

Participant ID	Age	Gender	Job Title	Role	Workshop Attended
P111	24	Female	Start-Up Founder	Role 1	Workshop 1
P112	40	Male	Conversational AI Consultant	Role 1	Workshop 1
P114	40	Male	Conversational AI Consultant	Role 1	Workshop 1
P115	25	Male	Conversational AI Consultant	Role 1	Workshop 1
P116	38	Female	Conversational Coaching Designer	Role 1	Workshop 1
P212	38	Male	Voice User Experience Designer	Role 1	Workshop 2
P213	25	Female	Conversation Designer	Role 1	Workshop 2
P215	40	Female	Conversational AI Consultant	Role 1	Workshop 2
P122	22	Female	Researcher	Role 2	Workshop 1
P123	24	Female	Start-Up Founder	Role 2	Workshop 1
P125	23	Female	PhD Student	Role 2	Workshop 1
P222	22	Female	Banking Design Engineer	Role 2	Workshop 2
P223	29	Female	Researcher	Role 2	Workshop 2

of the workshop. Recorded interviews were transcribed and underwent thematic analysis (Braun & Clarke, 2012). Two researchers coded the data separately and had a strong inter-coder reliability (Cohen's Kappa > 0.80), a third researcher was then involved in discussions on differing codes until a consensus was reached, resulting in two passes in total. This process resulted in several top-down and bottom-up codes (Table 5). The top-down codes (7) aligned with study objectives: "Activity 1", "Activity 2", "Activity 3", "Technical Utility", "Value Elicitation", "Toolkit Suggestions." Meanwhile, bottom-up codes (7) were created based on researchers' interpretation of the data (Braun & Clark, 2019): "Education or Mindset Change", "Increasing Practitioner Value-Sensitivity", "Collaboration", "Participation", "Using the Toolkit", "LLMs or Gen AI", "Types of Categories of Values."

5.2. Results

5.2.1. Activities

Codes relating to the three activities of the VSCA workshops, which align with the three steps in the VSCA framework, are collated under the "Activities" theme. Here, participants describe their experiences with the different activities and provide their perspectives, opinions, and other useful remarks.

5.2.1.1. Activity 1 - Value scenarios with technical considerations. Participants discussed activity 1 as a way of "aligning expectations" [P112], providing "rationale" [P111] for creating a CA and preventing "feature creep" [P115]. They appreciated the **equal importance placed on technical and value-based aspects** [P116], forming links between use-cases and values [P123], gaining the "vocabulary" [P123] and "jargon" [P125] to create a shared "mental image of what [they] want" [P125].

5.2.1.2. Activity 2 - Deriving and prioritising value requirements. Regarding activity 2, participants viewed the canvas as **a tool to systematically consider and discuss values**. They mentioned the visual nature of the canvas helping them to guide discussions [P213] and deeper reflections [P222], especially for sensitive topics such as values [P116]. **More examples were desired** [P115, P222, P116, P223, P214], but providing too many example was seen as "leading" [P116] and "restrictive" [P115]. Participants mentioned the activity helped **stay focused on a project's scope** [P215] and **reach shared definitions** [P115]. They mentioned the canvas also acts as documentation to revisit the definitions throughout the process to check if they have changed [P116].

5.2.1.3. Activity 3 - Value prototyping for embodiment. Activity 3 allowed participants to **design flowcharts and conversations with values in mind**. Intended values influenced participants' flow charts [P123, P223, P222], conversation content [P115, P116, P117], CA interactions (e.g., showing a typing status or using filler words) [P212, P213, P214], the CA's tone and personality [P212, P214, P111], and the CA's functionality (e.g., allowing users to invest in entities aligned with their values) [P214, P211]. Participants appreciated **integrating previous steps** [P115] to "practically combine what [they've] done in the first steps" [P215], and were "surprised by how it all came together at the end" [P116].

5.2.2. Benefits & praise

The framework and toolkit were observed to provide benefits to workshop participants themselves, many of whom directly offered praise. These comments and ideas are discussed under the codes of "Education or Mindset Change" and "Increasing Practitioners' Value Sensitivity."

5.2.2.1. Education or mindset change. Several participants described changes in their mindset or new information or ways of thinking that they have learned due to using the VSCA framework and toolkit. These are summarised below:

- The importance of communicating with end-users [P212, P213],
- Considering values when designing technology [P213],
- Considering both technical and value-based aspects [P116, P213],
- Differentiating between user needs or functionalities and user values [P223],
- Breaking down and structuring work with values [P222, P223]

5.2.2.2. Increasing Practitioner Value sensitivity. Participants described the framework's process as a "journey" [P114], with equally important parts that flow and combine well to foreground values [P114, P116, P125, P222, P123, P122]. They **emerged with heightened sensitivity to human values**, described as: "open[ing their] mind to more values" [P111], "filling in the 10% that a nice, empathetic designer doesn't already know" [P212], having "value glasses on" and "awareness/alertness for values" [P116].

5.2.3. Interactions

One of the cornerstones of the VSCA framework and toolkit is co-design and the support of interdisciplinary collaboration. Accordingly, the codes "Collaboration" and "Participation" focus on these aspects.

5.2.3.1. Collaboration. Overall, participants felt the toolkit **helped establish** "common ground" [P122] and a shared "mental image" [P125] between CA users and creators,

Table 5. Themes and codes created by the researchers based on workshop recording transcripts and post-workshop interview transcripts following thematic analysis. The number of transcripts in which each code appears and representative quotes are also provided.

Theme	Code	N (%)	Example quote
Activities	Activity 1	8 (31)	"I like that we were having a rationale behind how we're building AI conversational experiences. So you actually have a good thought of: 'OK, these are the stakeholders. What do they need? What type of features? What can we do about it?'. Like what are the technological means that could build this?".
	Activity 2	9 (35)	"The part with the values, with the round picture, with all the different values, there was something you know for the group to talk about and like really go through, 'OK, what do you think? Do you think this is relevant or not?'".
	Activity 3	17 (65)	"I think even just building the conversation with values taken into account, it's probably the moment where it's practically combined what we've done in the first steps and then we come to this practical exercise and I think it's really useful, but I think it's like many things if you just learn the theory it doesn't really settle with you and when you start applying it, then you really understand how you can incorporate it his into your actual work."
Benefits & Praise	Education or Mindset Change.	7 (27)	"I think the first takeaway would be that it makes sense to think about values beforehand compared to sometimes you're just pulling them out to justify something to a client".
	Increasing Practitioner Value Sensitivity.	16 (62)	"But I think through the workshop I gained that awareness more in the sense of in the process of designing the entire conversational process, so I will definitely take away that one. I'm gonna be working on stuff like that to visualize the values that I really wanna implement here and so that is definitely a big gain and value awareness."
Interactions	Collaboration	13 (50)	"I had some really interesting discussions and I really see the value, in having the two perspectives start at different ends and then coming together."
	Participation	15 (58)	"I think it really gives language to talk about and talk about why something's don't work in some context."
	Technical Utility	14 (54)	"You really understand how you can incorporate this into your actual work."
	Toolkit Suggestions	15 (58)	"You have examples in the conversation patterns or discussion prompts, and that's helpful. In the emotional one I missed examples like I don't even know like some use case examples or like some either matched matching conversations with different values involved or I don't know like checking that abstract form of values into some more functional examples for different use cases."
Toolkit Goals	Using the Toolkit	12 (46)	"I think it's really good to use it at the start because again, there's some decisions about like how you design whatever you build and like if you put it in the correct perspective from the beginning, it just helps you avoid, you know useless iterations."
	Value Elicitation	9 (35)	"The ideation of the values were based on my own values, but then once they were kind of like the group agreement, this is then you kind of feel a bit more distance from whatever you said because it feels like it's not just you saying it, it gets moulded by other people as well."
	Value Embodiment	8 (31)	"I think it kind of really helped ground the conversation we were having and personalize it a bit like we were forced to integrate very literally into the problem areas we were discussing clearly, I think it was also good to be thinking about the positive capabilities of these technologies, I guess the concept of of what a successful iteration of this kind of design would look like was really exciting".
	LLMs or Gen AI	9 (35)	"I was thinking of how LLM's don't usually say that they don't know how to answer a question they don't know the answer to, and they usually just give a wrong answer instead of saying 'ohh actually I'll just forward you to a person that knows this' and realizing early on that it's not in its capabilities of answering it. I was just going back to our ideas and the previous exercises where we were saying you're just like for dignity, I want to be treated like a person and want to be taken seriously and not waste time, so if it's actually outside of the capabilities just immediately say this is not something for me to answer."
Values and CAs	Types or Categories of Values.	11 (42)	"Things like usability or efficiency or safety, they're always there. And we always kind of think about them, but then there may be some things that like if you start focusing more on these values, you just figure out that you know the flow of conversations would be different."

aligning expectations. CA creators appreciated collaborating with users instead of "sitting in offices … and just do this on [their] own" [P114], while users learned creators' perspectives [P123, P111, P125]. Diverse and interdisciplinary experiences also enriched discussions [P221, P213, P114, P111, P115, P116]. **Participants had mixed feelings on group compositions.** Some preferred having CA users together initially for approachability [P123, P125], and having CA creators together for productive conversations [P116] before combining both. Others wanted more "cross pollination" [P115] between CA users and CA creators to mix perspectives [P215], and to help CA users be more effective [P222].

5.2.3.2. Participation. The framework gave participants **a sense of empowerment**. For creators, it provided "language" [P215] and "justification" [P213] to avoid client-based "feature creep" [P115], focus on core values [P112], and communicate effectively with non-technical stakeholders [P214]. For CA users, it gave them vocabulary and awareness to participate meaningfully [P123, P223, P125]. Some found it **daunting to speak up** due to age [P222] or expertise, feeling that CA users and creators had "[un]equal authority" [P115]. Nevertheless, CA users felt empowered to hold CA creators accountable after expressing their values and outlining how to embody them [P122, P123, P125].

5.2.4. Toolkit goals

This theme revolves around the objectives of the VSCA framework and toolkit. The codes of "Technical Utility", "Value Elicitation", and "Value Embodiment" relate to the core components of the framework, while the code "Toolkit Suggestions" expresses participants' suggestions for changes and improvements, and the code "Using the Toolkit" explores participants' comments on when they would use the toolkit throughout the CA design process.

5.2.4.1. Technical utility. Technical utility, in the context of this study, relates to supporting processes and creating outcomes that are useful from a technical perspective, and are able to drive implementation, without needing additional overhead or extra work. Participants described the framework and toolkit **facilitating technical utility** in several ways:

- Breaking down "implicit" values so that "you really understand how you can incorporate this into your actual work" [P215],
- Providing a "modular" [P112] toolbox that "you can adapt to certain situations or add on or remove some things" as needed [P111],
- Bringing together everything practically into a sample conversation [P115, P215, P116],
- Providing documentation to refer back to and offer accountability [P116, P123, P122],
- Providing "guardrails" [P115, P116], "tracks" [P115] and "skeletons" [P125] to structure conversations,
- Providing checks at the end to ensure values have been incorporated and respected [P213],
- Facilitating various collaborations with different stakeholders across the design process: "So that's how I would have used these three types of activities. It's almost like #1 is together with my development team, #2 is with me doing that experience design and #3 is with the customer or with the user" [P111],
- Combining with and "enrich[ing]" existing methodologies [P112],
- Prioritising specific values and addressing trade-offs [P123],
- Getting diverse stakeholder inputs for one-person teams [P213].

5.2.4.2. Toolkit suggestions. Participants' suggestions largely revolved around **providing more examples** for using the toolkit without facilitation. They had mixed feelings about adding more examples of values in Activity 2 and segregating creators and users. CA users recommended adding space for "anti-values" [P123] that are not desired and "anti-users" [P215] with malicious intentions. One participant suggested **using other value taxonomies** if desired [P115]. For Activity 3, participants requested **spaces to add values on flowcharts**, as most did so with sticky notes. A few suggested **including persona templates** to attach values to [P112, P114], and supporting other stakeholders besides CA creators and users, such as CA managers and executives [P114].

5.2.4.3. Using the toolkit. The majority of participants envisioned **using the framework and toolkit starting at the beginning of a project** [P114, P213, P215]. Additionally, several participants discussed **using it iteratively throughout the CA design process** [P112, P215]. One participant especially appreciated **the toolkit being usable during the** "middle" **implementation stages of the design process** [P115].

5.2.4.4. Value elicitation. As mentioned, CA creators felt the framework helped reach "the 10% that a nice empathetic designer doesn't already" [P114], while CA users felt their values were "at the forefront" before "get[ting] molded by other people" [P122]. The framework helped identify "anti-values" [P123] and surface deeper values not normally considered [P215, P223, P222]. It aided prioritising values and considering trade-offs across stakeholders [P213]. Two participants mentioned **Miro's interactivity and multi-modality them helped work with values more effectively** [P116, P213].

5.2.5. Values and CAs

Beyond the VSCA framework and toolkits, the study produced many valuable insights regarding working with human values when creating CAs. These are discussed using the codes "LLMs (Large Language Models) and Gen AI (Generative AI)", and "Types or Categories of Values."

5.2.5.1. LLMs or Gen AI. Regarding LLM-powered CAs, participants agreed that **more values become critical** as LLMs enable sophisticated use-cases [P222]. One participant mentioned **thinking about human values can justify whether to use an LLM-powered CA**: "I'm not sure that LLMs can just cover it all [referring to the intended values] … because it might be easier to do it with … classic approaches. So it really gives language to talk about why something's don't work in some context" [P215]. Another suggested **developing best practices for training LLM bots out-of-the-box** using commonly applicable values [P212]. Participants also discussed how **LLM-powered CAs bring new values into play**: "I was thinking there was this Chevrolet LLM bot … people were asking it to do things completely unrelated to its use case and it was doing them … Relevance could be also a value that you tend to forget … just making sure it's sticking to stuff that it's related to the brand, especially with LLMs now" [P213].

5.2.5.2. Types or categories of values. Throughout the workshops and interviews, participants explicitly and implicitly created their own groups or categories of values. These groups were based on:

- "Intuitive" [P215] values versus non-intuitive ones. Intuitive values were always functional values which were framed as an "expected given" [P221] or "default" [P215], versus epistemic or social values [P213]: "I think that's huge win from a toolkit like this is because we get so lost into these functional values. But then you forget

about epistemic like all these learning curiosity, self-determination, agency, it was quite interesting" [P111],
- Based on different use-cases [P112],
- Based on user groups and demographics [P114, P125],
- Based on the role of the CA (e.g., accountant versus advisor) [P125],
- Values coming from a human versus an AI [P213]: "I'm also thinking of it as like an interpersonal value that seems like it respects your privacy? But I don't think that's really how it works with the AI, it's probably like a functional thing" [P223],
- Which stakeholder the values belong to [P215],
- Different user states [P214].

5.2.6. Summary of results

Workshop transcripts and interviews yielded several insights: Activity 1 helped align expectations across CA creators and CA users, define a user-centered scope, and balance between technical and value-based aspects. Activity 2 facilitated systematic value consideration, and created shared value definitions and hierarchies. Participants had mixed views on the amount of examples provided. Activity 3 integrated previous activities' outcomes through flowcharts and conversations, incorporating intended values into numerous design decisions. The framework and toolkit improved participants' value sensitivity and collaboration. It enabled them to establish common ground and empowered CA creators to focus on core values and CA users to express their values and hold CA creators accountable. Participants had mixed feelings regarding segregating CA creators and CA users during activities.

Participants attested to the technical utility of the outcomes produced, envisioning iterative use of the framework throughout the design process and at the end of it, to verify translation of intended values to embodied values. CA users reported expressing their values effectively, while CA creators successfully described how different values can be embodied in prototypes created. The multi-modality offered by the toolkit facilitated working with values more effectively. Participants had several suggestions for improving the toolkit. Finally, participants discussed LLM-powered CAs' impact on values, and vice versa, assessing LLM compatibility with intended values. They also created their own groups and taxonomies of values that focused on other contextual aspects beyond the nature of the values themselves.

6. Comparative survey study - Assessing the value embodiment of conversational agent prototypes (framework outcomes)

6.1. Methodology

6.1.1. Procedure

A comparative study was conducted to compare between two VSCA scripts and two control scripts. A survey was created using Qualtrics and consisted of three questions. Participants were randomly assigned to either see the first VSCA script and control script (general budgeting advice), or the second VSCA script and control script (advice for saving toward a specific goal). The order of script presentation was also randomized to counterbalance learning effects. Participants viewed one script, answered three questions, then viewed the other script and answered the same three questions again.

Assessing the value-sensitivity of a technology is challenging, due to the abstract nature of values (as acknowledged by the creators of VSD (Friedman et al., 2021)). Related approaches include AI creators self-assessing how their system fosters AI ethics values (Yurrita et al., 2022; Hallensleben & Hustedt, 2020), or verifying value-sensitivity through an ethics matrix (Mepham et al., 2006) filled by researchers based on discussions with target users (Strikwerda et al., 2022). However, these methods can lack empirical rigor and be susceptible to researcher subjectivity. Recent work has distinguished between evaluating any Responsible AI tool based on its "usability" versus its "effectiveness in impacting AI development" (Berman et al., 2024, p. 1). Berman et al. (2024)'s review of evaluation methods highlight an overwhelming focus on usability, while neglecting to evaluate the impact that tool usage has on the resulting AI-powered system. Accordingly, the survey focuses on understanding the framework and toolkit's effectiveness by operationalising van de Poel (2020)'s model of translating Intended Values (captured during the workshops) into Embodied Values in the prototypes (as perceived by target users, i.e., bank clients). In line with this, we define three metrics for evaluating effectiveness in terms of our initial goals for the framework and toolkit (Berman et al., 2024). These three metrics correspond with the three survey questions in Table 6.

6.1.2. Participants

The survey was posted on Prolific and completed by 451 participants who were paid £9 per hour. Participants were screened for having a bank account, as this meant having experience with banking, making them potential target users for a banking CA. They were also screened for having English as their primary language, given that the survey was in English. Leveraging target end-users to evaluate the effectiveness of AI tools is a recommended practice (Berman et al., 2024). 10 participants were used as a pilot to identify any major issues with survey design, question clarity, technical functionality, and overall flow. Their results were then disregarded as changes were subsequently made to the survey. Additionally, 35 participants' answers were removed for providing incomplete answers, where one or more questions were left blank. Afterwards, 406 participants were included. For the first set of scripts (general budgeting), 187 participants saw the VSCA and Control scripts in different orders. For the second set (saving for a specific goal), 219 participants saw the VSCA and Control scripts, with different viewing orders. See Figures 10 and 11 for details.

Table 6. The three survey questions included, as well as their objectives and corresponding metrics probed.

Q#	Question wording	Question format	Question objective	Metric	Notes
Q1	What values do you think are being respected or followed by the chatbot in the conversation you just read? If you cannot think of any values, please type anything in the box titled 'I don't know'.	Text Box	To understand the proportion that intended values represent out of all perceived embodied values.	Value Efficiency	The survey begins by explaining what values are and provides examples of a few values.
Q2	Based on the conversation above, to what extent do you feel the chatbot aligns with the following values? You can find explanations of each value above this question.	Several Likert Scales	To understand the extent to which intended values are translated to perceived embodied values.	Value Translation	Participants are given a list of the intended values from the corresponding workshop and their definitions. The second script is only compared to the intended values of workshop 2 as values across a project are dynamic and expected to change and evolve and so the most recent iteration of values was used (Manders-Huits, 2011) (Jakesch et al., 2022).
Q3	Think about other values that might matter to you in the context of banking. Overall, to what extent do you feel the values being respected or followed by the chatbot in the conversation align with your values as a bank client?.	One Likert Scale	To understand the extent to which perceived embodied values are compatible with target users' values.	Value Compatibility	N/A

6.1.3. Data collection & analysis

Data collection accounted for the minimum required sample size of 344 responses, which was obtained by conducting a power analysis with a medium expected effect size (0.5), a significance level (α) of 0.01, and a power level ($1 - \beta$) of 90% for two groups with a $\sim$ 50% split (Curtis et al., 2015; Balkin & Sheperis, 2011). Extra responses were also collected to account for any drop-outs and incomplete responses, given the complexity of the survey. Each question was analyzed separately, with results segregated by script version (version 1 or version 2). For Q2 and Q3, results were further segregated by viewing order due to its significant impact on participants' ratings.

6.1.3.1. Question 1 - Value efficiency. Question 1 resulted in four lists of perceived embodied values - one for each script. The unique values for the first VSCA script and Control script were aggregated and mapped (based on paraphrasing) onto the intended values emerging from the first workshop by the consensus of two researchers, building on similar value mapping techniques from prior works (Strikwerda et al., 2022; Sadek, Constantinides, et al., 2024). The same procedure was conducted with the second VSCA script and control script. Afterwards, the previous lists of values for each script were replaced with lists of binary "yes" or "no" depending on whether the value in each cell had been mapped under one of the intended values or not. For each pair of scripts separately, these binary lists were converted into a contingency table. McNemar's test with Phi Coefficient determined significance and effect size.

6.1.3.2. Question 2 - Value translation & question 3 - Value compatibility. Questions 2 & 3 underwent similar analysis processes. Firstly, exploratory analysis revealed non-parametric data and viewing order significance (Shapiro-Wilk, Mann Whitney U). Bonferroni Correction was then applied. Finally, Wilcoxon Rank Sum was used to compare the paired data (a control script and VSCA script rated by the same person). For Q2, each person's average across Likert scales (10 for version 1, 9 for version 2) represented overall alignment between intended and perceived values. For Q3, direct Likert rating was used. Outliers were trimmed as results did not differ significantly.

6.2. Results

6.2.1. Question 1 - Value efficiency

As shown in Figure 12, embodied values mapped better to intended values for VSCA scripts compared to Control scripts in Empathy & Understanding, Playfulness & Friendliness, Emotional Support, Dignity, and Safety & Security. These are largely more human-centred, non-functional values (Fuchsberger et al., 2012). Control scripts performed better in Building Trust and Transparency & Clarity. Workshop participants had defined the intended value of "Building Trust" as "The CA builds the user's trust in it by not asking for any unnecessary information, explaining why and how things are happening, and so on", implicating related values around transparency and efficiency, which are more functional values (Fuchsberger et al., 2012). These results suggest that the VSCA scripts perform better with relation to non-functional values, as CA creators might already be more familiar with functional values (Bingley et al., 2023; Sanderson et al., 2023). Table 7 summarizes differences the proportion that intended values represent out of all perceived embodied values. Both versions showed significant differences, with VSCA scripts

Time Taken (seconds)	Maximum	Minimum	Average	Std
	3483	187	873.0	490.7

Age (Years)	Maximum	Minimum	Average	Std
	70	18	29.9	9.2

Sex (Count)	Male	Female
	191	215

Ethnicity (Count)	Asian	White	Black	Mixed	Other	Data Unavailable
	15	261	97	23	9	1

Employment Status (Count)	Student	Full-Time	Part-Time	Unemployed	Not in paid work	Data Unavailable	Other
	143	159	31	31	8	28	6

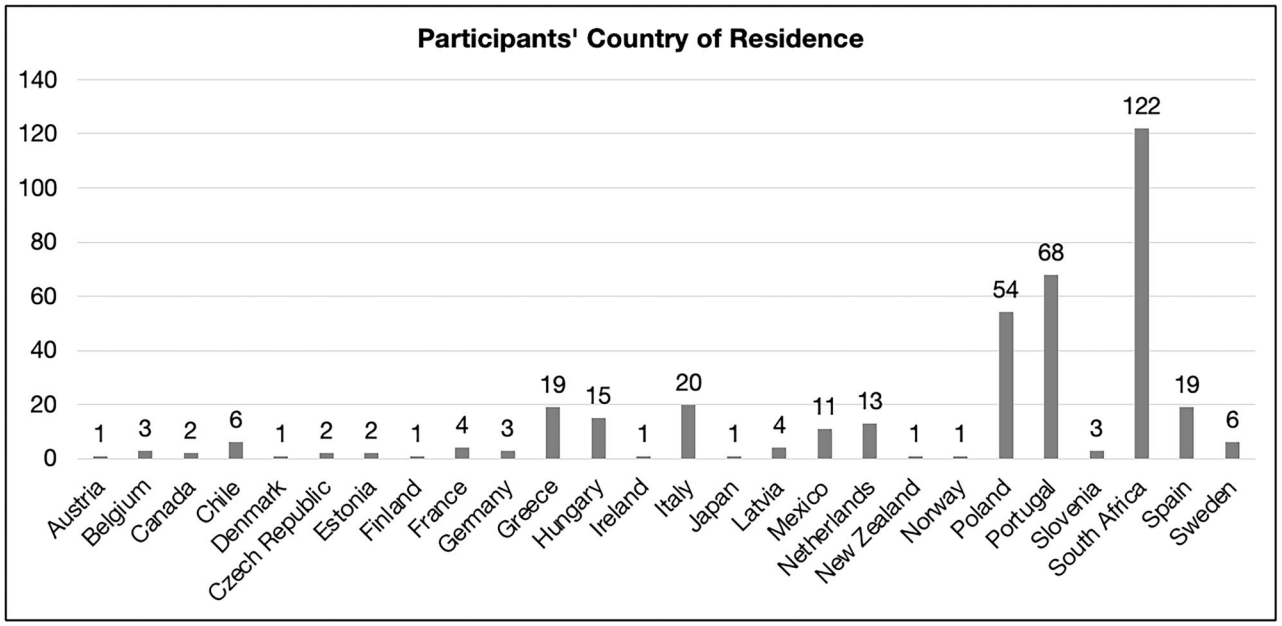


Figure 10. A summary of various survey participants' demographics including age, gender, ethnicity, country of residence, employment status, and time taken to complete the survey.

assigned more perceived values overall. The second VSCA script's effect size was slightly larger than the first one.

6.2.2. Question 2 - value translation & Question 3 - value compatibility

VSCA scripts rated significantly higher than Control scripts, independently of viewing order. Control script ratings decreased if VSCA scripts were seen first, while VSCA script ratings remained consistent. Question 3 results followed Question 2 patterns as shown in Figure 13. Value Translation and Value Compatibility differed significantly between VSCA and Control scripts across versions and viewing orders, with no consistent effect size difference across versions. These results are shown in Table 8.

7. Discussion

7.1. Theoretical implications - Guidelines for practically supporting value embodiment in conversational agents

Based on the findings of this study, we offer a set of guidelines for practically supporting value embodiment and foregrounding human values when designing CAs. These guidelines are intended for practitioners who fit under the group of CA creators (e.g., designers, developers, etc.) and are summarised and expanded upon below.

1. **Value-Based Scoping:** Align stakeholder expectations using shared and agreed-upon intended values to avoid feature creep and foreground value considerations.

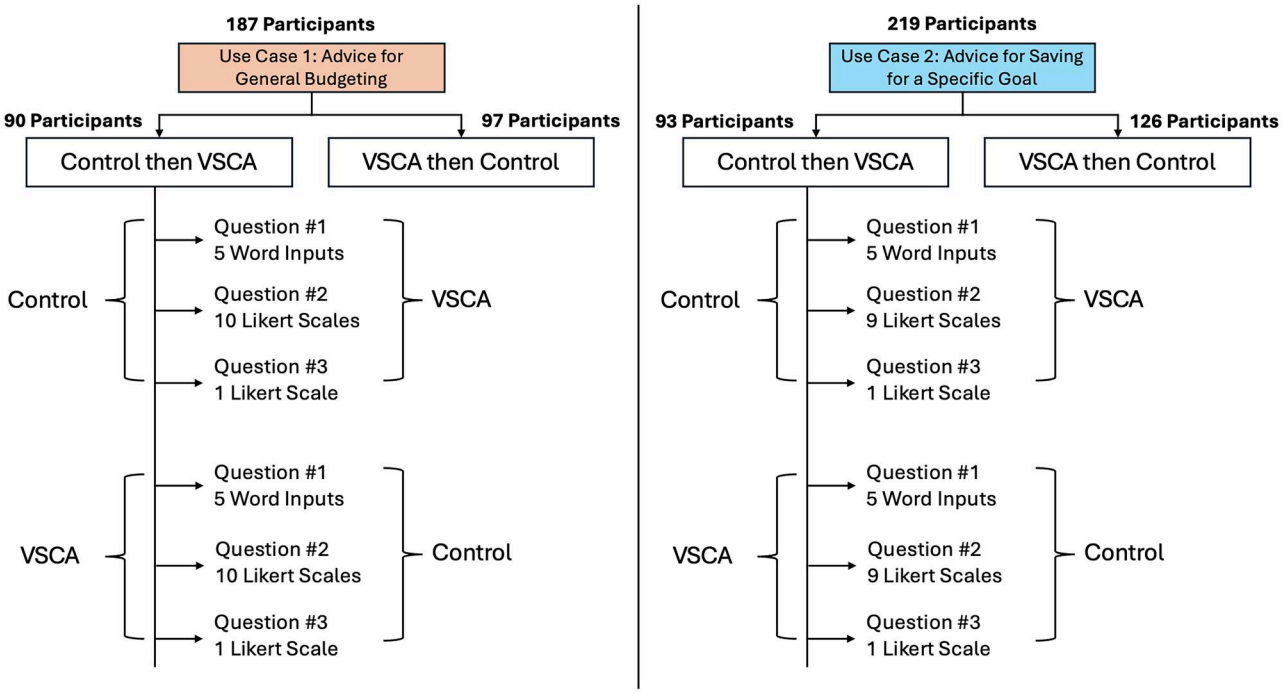


Figure 11. A flowchart outlining the structure of the survey. An individual participant would be randomly assigned one of four possibilities.

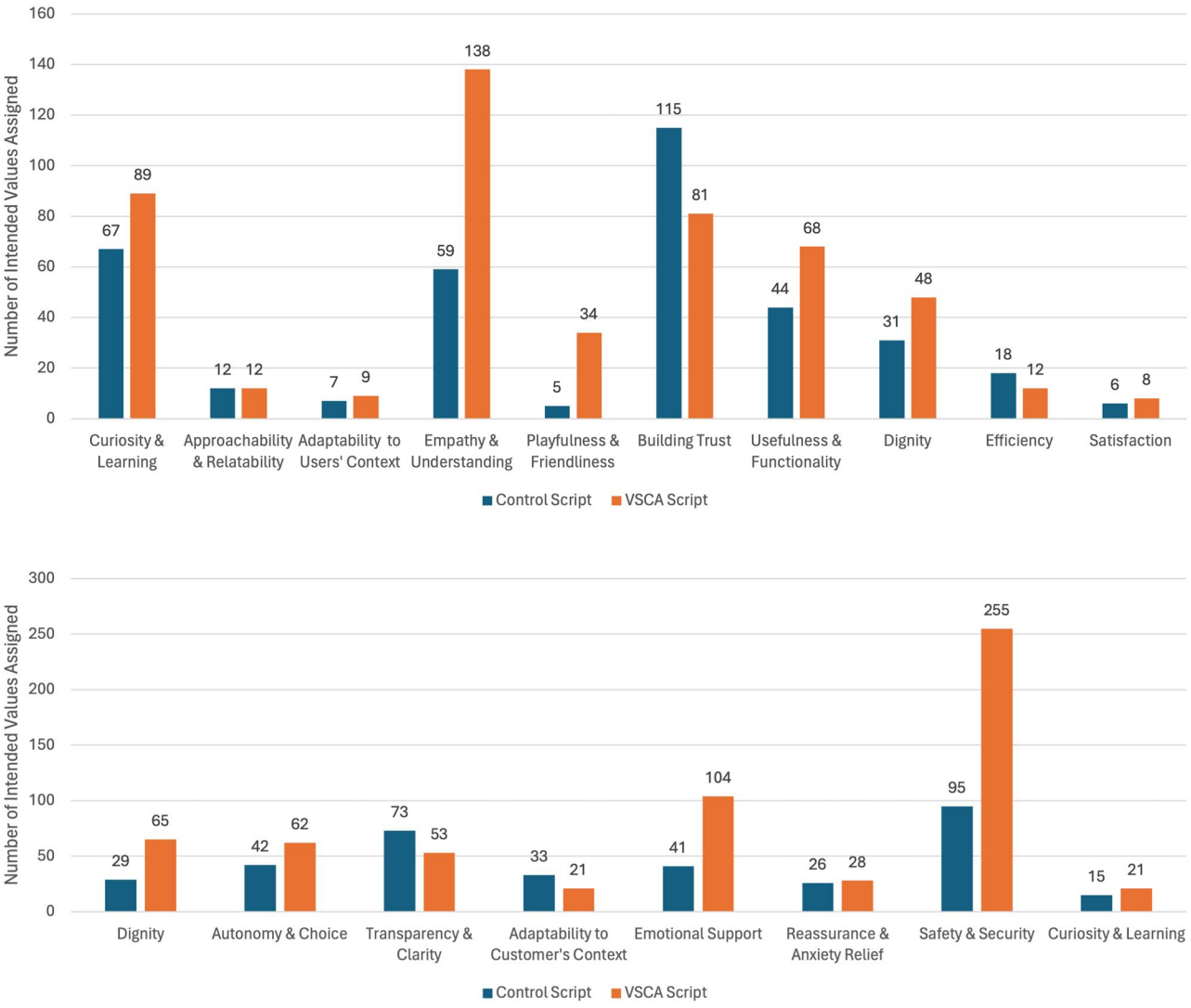


Figure 12. The number of perceived embodied values assigned by survey participants that map to the original intended values in the first version of VSCA and control scripts (top) and the second version of VSCA and control scripts (bottom).

Table 7. Question 1's results showing the numbers of perceived embodied values that mapped to intended values across both versions of VSCA and control scripts as well as the results of testing for significant differences across VSCA and control scripts.

Perceived embodied value mapped to an intended value?.	VSCA Script	Control script	Test statistic	*p*-value	Effect size
Version 1					
Yes	503	370	172.40	<0.0025[a]	0.42
No	88	29			
Version 2					
Yes	609	354	223.16	<0.0025	0.46
No	52	46			

[a]Threshold based on Bonferroni Correction with a desired significance level of 0.01.

Previous studies highlighted difficulties with unaligned stakeholder expectations (Vogelsang & Borg, 2019; Zdanowska & Taylor, 2022; Kross & S, 2022). While experts recommended aligning expectations early regarding AI capabilities (Luger & Sellen, 2016; Amershi et al., 2019), we additionally recommend using shared intended values to create expectations around supported use-cases and underlying relevant values (as done in Activity 1 of the VSCA framework).

2. **Prompting & Scaffolding Value Discussions:** Provide prompts and examples to support value discussions. Participants stated visual prompts strongly helped to

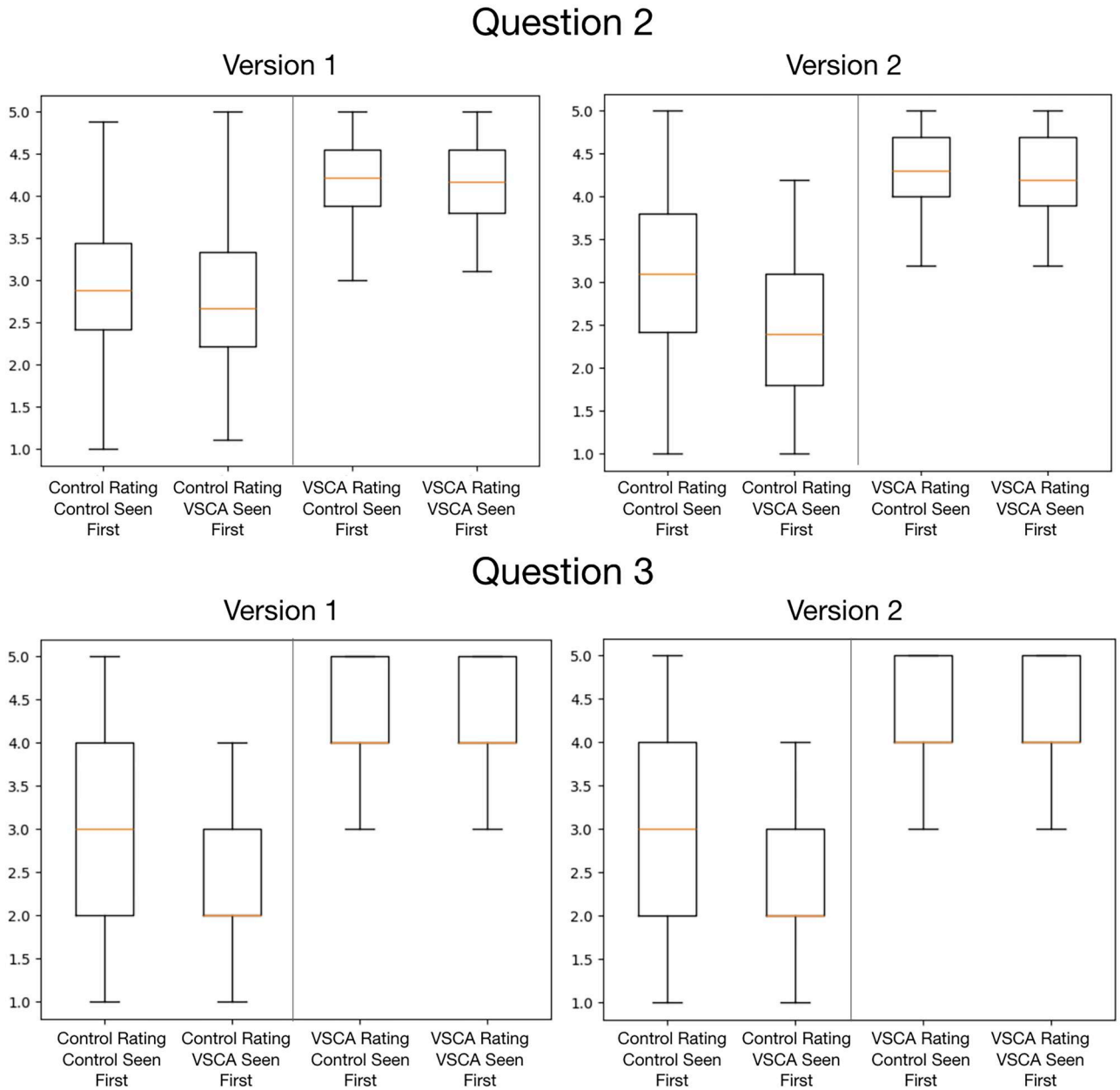


Figure 13. Boxplots showing survey participants' Likert scale ratings from 1-5 regarding (i) the extent to which they perceive intended values as being embodied by the VSCA and control scripts (top) and (ii) the extent to which they perceive the embodied values as being compatible with their own values (bottom).

Table 8. A Table summarising all the results from questions 2 and 3 including the mean and standard deviation of participants' ratings across different versions, viewing orders, and scripts, as well as the results of significance testing.

Version	Viewing order	Script type	Mean	Std	Test Statistic	*p*-value	Effect Size	Outcome
Question 2								
Version 1	Control	Control Script	3.104	0.952	−8.661	<0.0025	−1.810	Significant
	First	VSCA Script	4.339	0.458				
	VSCA	Control Script	2.487	0.830	−11.230	<0.0025	−3.063	Significant
	First	VSCA Script	4.265	0.488				
Version 2	Control	Control Script	2.912	0.611	−9.713	<0.0025	−2.388	Significant
	First	VSCA Script	4.20	0.227				
	VSCA	Control Script	2.875	0.525	−10.116	<0.0025	−1.920	Significant
	First	VSCA Script	4.18	0.408				
Question 3								
Version 1	Control	Control Script	2.989	1.022	−7.781	<0.0025	−1.560	Significant
	First	VSCA Script	4.305	0.642				
	VSCA	Control Script	2.379	0.976	−9.772	<0.0025	−2.152	Significant
	First	VSCA Script	4.158	0.704				
Version 2	Control	Control Script	2.968	1.146	−7.522	<0.0025	−1.365	Significant
	First	VSCA Script	4.310	0.653				
	VSCA	Control Script	2.400	0.985	−11.405	<0.0025	−2.395	Significant
	First	VSCA Script	4.250	0.579				

discuss and systematically elicit intended values. Open-ended prompts have been found to improve stakeholder collaboration (Sadek, Constantinides, et al., 2024) and facilitate non-technical stakeholder participation (Wong et al., 2022) in general, and our results show this extends to value-based considerations as well. Participants requested more examples due to unfamiliarity with working with values, but found it difficult to think beyond provided examples. This "fixation effect" (Jansson & Smith, 1991) or "anchoring bias" (Furnham & Boo, 2011) is well documented across a number of domains. Instructing participants to avoid repetition, using abstract examples (Ezzat et al., 2020), and showing examples later on in the activity (Sio et al., 2015) can reduce such fixation.

3. **Documenting Value-based Outcomes:** Create accessible documentation to define intended values, update them as they evolve, and verify their translation to embodied values. Boundary objects enable transparency and accountability (Custis, 2021; Go FAIR, 2021), addressing the lack of suitable documentation in AI systems that are usually only made *by* practitioners *for* stakeholders (Piorkowski et al., 2021). Yildirim et al. (2023) recommend creating boundary objects usable by different stakeholders, this is still found to be lacking for CAs (Heo & Lee, 2023). While previous works have established the importance of documentation in design, AI, and CAs, we further establish its importance for documenting intended values, their definitions, their evolution, and their translation to embodied values, as recent work has begun to acknowledge and accommodate for (IEEE 7000, 7000, 2021).
4. **Establishing Value Hierarchies:** Accommodate value trade-offs and prioritise values by creating hierarchies of intended values based on context. Activity 2 prioritises values as "critical", "important", and "nice to have." This categorisation establishes hierarchies to rank values, resolve tensions (Manders-Huits, 2011) and conflicts (Palmer & Schwan, 2024), de-bias values (Umbrello, 2018), and guide technical practitioners. Our study showed that the same value can have different importance for stakeholders, aligning with recent work (Jakesch et al., 2022). Consensus-building activities between stakeholders (e.g., Value Dams and Flows used in Activity 2) are therefore critical when considering values.
5. **Moving Beyond Predominantly Functional Values:** AI practitioners (Vogelsang & Borg, 2019) and CA practitioners (Sadek et al., 2023b) tend to focus on technical and functional/profit-oriented values, and those related to regulatory compliance (Morley et al., 2023), overlooking epistemic, social, interpersonal, conditional, and emotional values. AI practitioners find human-centred values difficult to work with, failing to see their applicability (Sanderson et al., 2023). Previous work has highlighted that empathy and related values are expected by humans during communication (Preston & de Waal, 2002) and is especially valuable when communicating with CAs, particularly in sensitive domains such as healthcare (Li et al., 2024). VSCA's results show that prioritising different categories of values concurrently is possible and does not impact performance across the different categories.
6. **Enabling Value Prototyping:** Recognise and utilise the multiplicity of ways that values can be embedded and incorporated into CAs, and practically integrate value-based outcomes into prototypes. Our study has highlighted that values can be integrated into CAs' interactions, tone and personality, and functionality; not just in their content. While prototyping is a foundational design activity and a core part of the CA design process, we also recommend using it as a checkpoint to ensure intended values are successfully propagated into embodied values (van de Poel, 2020).
7. **Fostering a Culture of Value-Sensitivity among Team Members:** Encourage team members to learn about human values' importance, widening technically-

focused practitioners' horizons (Morley et al., 2023; Dindler et al., 2022). Practitioners are often unaware of implicit values they embed in their work, which can conflict with intended values (Fest et al., 2023). The neglect of value-related aspects among practitioners (Morley et al., 2023; Dindler et al., 2022) leads to responsibility shifting (Mitchell, 2020; Morley et al., 2023) and a focus on passive compliance (Sadek, Kallina, et al., 2024). Fostering value-sensitivity builds empathy, avoiding these unwanted behaviours (Sadek, Constantinides, et al., 2024). More in-situ interventions that integrate with their workflows can be effective (Wang et al., 2024), given practitioners' views that ethical aspects lie outside traditional computer science (Darling-Wolf & Patitsas, 2024).

8. **Considering Different Group Compositions for Value-Based Collaborative Activities:** Decide on the most appropriate group composition given the benefits and drawbacks of segregating or mixing different stakeholders (e.g., CA creators and CA users or technical and non-technical stakeholders). While co-design advocates designer-user collaboration (Sanders & Stappers, 2008), segregating CA creators and CA users had advantages when working with personal, sensitive values. It allowed CA users to ease into participation, and allowed CA creators to have productive conversations before combining. However, several participants wanted more inter-stakeholder collaboration. This was the case across other studies involving AI practitioners as well (Alspaugh et al., 2018; Zdanowska & Taylor, 2022).
9. **Recognising Stakeholders' Unequal Authority**: Encourage stakeholder empowerment and meaningful participation by actively evening the playing field across stakeholders. The "asymmetry" in power between AI users and AI creators was mentioned by two CA users and is a well-documented issue when co-designing AI (Dobbe et al., 2021; Deng et al., 2023; Cooper & Zafiroglu, 2024). The extent to which participants engage in discussion and sharing strongly affects the collaboration's end-result and group effectiveness (Steen, 2013; Bratteteig & Wagner, 2016). Responsible AI tools often ignore team dynamics (Berman et al., 2024). Alternatively, when using the VSCA framework, several participants mentioned being able to use it as a tool to "push back" (against clients, developers, etc.) and justify their choices. Techniques such as segregating CA users and CA creators at first, prompting CA creators to use more accessible language (Yang et al., 2018; Luria, 2018; Alves-Oliveira et al., 2021), improving CA users' AI literacy (Smith et al., 2023; Cooper & Zafiroglu, 2024), and emphasising that CA users are participating as "experts of their experiences" (Sanders & Stappers, 2008; Vines et al., 2013), have proven effective in empowering CA users to participate meaningfully.
10. **Embracing Modularity to Extend Further into the Design Lifecycle:** Instead of rigidly following one framework or methodology from start to finish, or not following one at all (Sadek et al., 2023b), embrace modularity and pick-and-choose tools allowing active value consideration and propagation beyond early design phases. Currently, the majority of CA co-design (Sadek et al., 2023a), and collaboration with non-technical stakeholders in AI in general (Piorkowski et al., 2021), occurs during the early stages of the design process for scoping and ideation, or during the final stages for testing and verification. Frameworks do not generally cover middle implementation stages, despite recommendations for doing so during CA design (Lewandowski et al., 2021, 2022). Accordingly, we recommend making use of a wide range of flexible and reusable co-design techniques that can be adapted and revisited throughout the CA design process, as highlighted by previous research (Sadek et al., 2023a).
11. **Leveraging Visual Elements and Multi-Modal Interactions:** Increase immersion and effectiveness when engaging with values by providing tangible and visceral ways of doing so. Given the abstract nature of values, participants appreciated the tangibility of interacting with sticky notes on a Miro board. When designing in general, tangible objects and toolkits have been found to be suitable for quick ideation and evaluation (Boujut & Blanco, 2003) and encourage participation and conversation (Redaelli & Carassa, 2018; Pennington, 2010), especially in interdisciplinary settings (Wilkes et al., 2016; Keys et al., 2017). In terms of interacting with CAs especially, recent work has found that providing the opportunity for direct manipulation improved users ability to prompt text-based CAs (Masson et al., 2024). Our study shows that these benefits might be especially pronounced when dealing with abstract topics such as values, whose consideration should therefore include multi-modal and tangible cues, prompts, aids, and interaction mechanisms when possible.
12. **Reflecting on LLMs' Impact on Values and Vice-Versa:** Understand whether an LLM-powered CA can deliver and respect intended values, and which values might need to be added, removed, or altered accordingly. Considering values becomes critical as LLM-powered CAs tackle more sophisticated tasks (Chen et al., 2023). Two important questions to consider are: (i) Can an LLM-powered CA deliver/respect intended values, or is another approach needed? and (ii) How will intended values be affected by LLM use; will new values be introduced, definitions updated, or existing values become irrelevant? In light of these questions and the fact that different LLMs have their own underlying values (McGrath, 2024), it is important to transparently highlight whether and how LLM-powered CAs will be utilised when using the VSCA framework or making any value-based considerations in general.
13. **Exploring Unfamiliar Values and Taxonomies:** To improve the inclusivity of CAs, explore taxonomies and values from various cultures, domains, languages, and groups. A participant suggested using value taxonomies beyond the Values in Action (ViA) approach (Fuchsberger et al., 2012). There are various value

taxonomies that exist (Schwartz et al., 2012; Yurrita et al., 2022) and likely other taxonomies in different cultures, domains and languages that we, as researchers, are not aware of. Encouraging colleagues, team members, and stakeholders to bring relevant taxonomies from their contexts can support this inclusivity. This can also be extremely beneficial in developing domain-specific CAs that are better value-aligned and less biased, as different domains, such as healthcare, can have specific professional codes of conduct, values, and ethical frameworks (Chow et al., 2023).

14. **Thinking About Value Sources and Value Owners:** Consider the origins and ownership of intended values and their implications. Participants noted that values might vary depending on whether they belong to a CA user, CA owner (e.g., the bank), or the CA itself. The value's "owner" could alter the value's definition, or the goals it drives. Recent work highlights this differentiation (e.g., "trustworthiness" in a CA vs. "trust" in a user) (Knowles et al., 2023). Prioritizing one value owner over another (e.g., researchers over participants) can significantly impact activity outcomes (Agnew et al., 2024).

7.2. Practical implications - Framework impact & evaluation

Overall, use of the framework yielded promising results in terms of technical utility and value elicitation (explored through qualitative analysis of workshop and interview transcripts) and value embodiment (explored through a comparative survey study of VSCA scripts and Control scripts). In terms of technical utility, CA creators expressed enthusiasm for using it in practice, leveraging outcomes for documentation and driving implementation with documented values/requirements. Meanwhile, CA users felt able to express values and confident in creators' understanding and incorporation of them.

VSCA scripts outperformed control co-design scripts, with significantly higher "value efficiency" (larger proportion of perceived embodied values were intended), more of which were human-centred values like empathy/dignity, and similar levels of functional values to the control scripts. VSCA scripts showed better value translation (intended values to perceived embodied values) and compatibility with users' values. The second version of the VSCA script was intended to evaluate the effect of iterating over the prototyping stage (Activity 3), undergoing two iterations over the two workshops. While it had improved value efficiency over the first version, suggesting iterating value-based prototypes could help, other factors require further investigation. The lack of improvement for value translation and compatibility could be due to a number of factors, such as participants of the second workshop being less skilled overall than the participants of the first workshop, or could be due to other confounding variables. Given the improvement for value efficiency and the general recommendations of co-design and VSD to iterate whenever possible and to include as many voices and stakeholders as possible, the VSCA framework will continue to recommend iteration until further evaluations have been carried out.

In terms of changes and updates to the framework and toolkit, segregation of CA creators and CA users, as well as striking the correct balance of examples and prompts provided, need further exploration before any changes can be made, which will be conducted in future work. Alternatively, changes will be made to: support using other value taxonomies; the inclusion of anti-values and anti-users; recommending tools for digitising prototyping and note-taking during Activity 3 (and potentially leveraging LLMs to rapidly prototype custom CAs (Kwon et al., 2024)); creating a persona and attaching intended values to it (recent work has began exploring the relationship between personas and LLMs (Ha et al., 2024)); and providing guidance and tools for evaluating the value embodiment of prototypes created while using the framework and toolkit, as opposed to at the end of the design process (Weidinger et al., 2023), based on the procedure of this study.

7.3. The emergence of large language models (LLMs)

Following the release of ChatGPT to the general public triggering a landslide of LLMs across industry (Toloka, 2023), and as academia enters its fifth wave of CA research pushing past the AI hype (Schöbel et al., 2024), the research community was quick to point out that creating human-centred Gen AI and LLMs, and particularly ones that align with human values, is as critical (if not more critical) than ever (Chen et al., 2023). Following this, a steady stream of research has shown that LLMs themselves have different sets of underlying values (McGrath, 2024; Yi et al., 2023; Pistilli et al., 2024). LLMs are neither omnipotent nor perfect (Rong & Li, 2023; Dutta & Chakraborty, 2023; Denning & Rousse, 2024; Videla, 2023) and their increased capabilities do not render efforts to align them with human values useless. Instead, they simply introduce new considerations when striving for value alignment.

As mentioned previously, practitioners who used the VSCA Framework and Toolkit agreed that they were still valuable and usable in the context of LLMs. However, future investigations are needed to understand the impact of Retrieval Augmented Generation (RAG), prompt engineering (Knoth et al., 2024), and other novel CA approaches on embedding values in LLM-powered CAs, and what this means in practice. Referring to the corresponding recommendation (#12) given earlier, a deeper understanding of whether an LLM-powered CA can deliver and respect intended values, and how values might need to be altered accordingly, is therefore needed from both researchers and practitioners.

As this space continues, and will continue, to rapidly evolve, the approach taken within the VSCA Framework and Toolkit of providing support for meta-level value considerations independently of technology or domain, while allowing space to mould both the process and its outcomes to suit specific contexts, future-proofs the intervention. It ensures the framework and toolkit remain relevant as they focus on engaging stakeholders and making sure practitioners are able and willing to embed values that should remain at the core of every technology. The focus on encouraging deliberate consideration from practitioners, and removing abstraction with regards to working with values, are re-usable and timeless.

7.4. Limitations & future work

The aim of this study was to introduce and evaluate the VSCA framework and toolkit. While the results for the framework and toolkit are promising and offered valuable insights into improving value embodiment within conversational agents in general, there are a number of limitations to consider.

Firstly, the choice of a between-subjects design for the creation of VSCA and control scripts. This was due to the limited time that industry practitioners could dedicate to the study (being only able to participate in one workshop). The focus on including industry practitioners was necessary given the focus on technical utility and practicality in real-world settings. As articulated by a recent empirical CA study, "this aligns with the push for "in the wild" HCI studies, in which the value of real-world observations is prioritised over highly-controllable laboratory settings." (Wester et al., 2024, p.11). A number of steps have been taken to improve comparability (having similar mean ages, gender compositions, areas of expertise, prior experience with interdisciplinary collaboration, and no prior experience with any framework related to human values), as well as using qualitative data to gain a more holistic understanding of the outcomes.

Secondly, researchers adapted second VSCA script based on both workshops' discussions/intended values, possibly introducing bias. However, given the similar results between the first and second versions, where the first version was created directly by workshop participants, this is unlikely the case. In all cases, the provision of two versions and two script viewing orders aims to provide more credibility to the effects of the framework and toolkit. Thirdly, while the sample size for the comparative study was large enough to draw statistical conclusions, the sample size for the workshops and interviews is more limited. Additionally, all participants were Western. Future work will aim to invite more participants, and especially more diverse participants, to use the framework and toolkit in order to gain a wider understanding of their impacts.

Thirdly, in terms of the evaluation conducted, the focus was on a finance/banking application. Future evaluation cycles will also need to span a number of different domains in order to verify the claims made about the generalisability of the framework and toolkit's impacts. Additionally, during the design workshops using the VSCA Framework and Toolkit, participants were instructed to take a modality-neutral approach when designing their prototypes. This meant that participants did not have to select a specific modality (e.g., text or voice) and take subsequent considerations specific to that modality. This decision was made to focus on the behaviour of the CA and assess the value-sensitivity of the CA's responses to chosen use-cases, without introducing other considerations that could impact embodied values. For example, perceived voice-based CA gender can impact user perception of "trust" (Tolmeijer et al., 2021), while perceived AI capabilities can impact user perception of "agency" (Schmitt et al., 2023). We also did not ask participants to focus on showcasing any specific AI capabilities, allowing them to brainstorm freely. Nevertheless, abstracting away modality and not highlighting recent advances in AI capabilities limited our ability to investigate the impact of these aspects on perceived values. Accordingly, future evaluation cycles will focus on investigating the incorporation of broader design decisions, including those that are modality-specific and capability-specific, to assess their impact on the outcomes of the framework and toolkit.

Finally, while not necessarily a limitation, it is important to note that the methods for evaluating value embodiment build upon van de Poel (2020)'s conceptual framework by empirically operationalising the framework to evaluate existing prototypes. Given the lack of evaluation methods within Value-sensitive Design (Friedman et al., 2021), this was deemed sufficient for this study.

Regarding future work, this will include updating the framework and toolkit as described in the previous section and collecting more data regarding their performance and impact. Specifically, the impact of the framework and toolkit on resulting CA prototypes' quality and user experience through established methods, such as the Conversational Agent Scale for User Experience (CASUX) (Faruk et al., 2024), should be explored, to compliment our understanding of their impacts on value-sensitivity. Another interesting avenue to explore is the incorporation of LLMs within the framework and toolkits themselves to aid with aspects such as developing the CA prototypes (Kwon et al., 2024) or underlying personas (Ha et al., 2024). This work also highlights a number of interesting avenues that we encourage researchers to explore, including:

- The effect of LLMs on values and vice-versa,
- The implications of using LLM back-end services through API calls to build CAs on the need for value-sensitive prompt engineering as opposed to value-sensitive conversation design,
- The effects of group structure and group composition when working with values,
- The reciprocal relationship between creator-user collaboration and articulating shared values, and how each one can enhance the other,
- How interventions and mechanisms can be created for ensuring embodied values are translated into realised values,
- The effects of providing examples and prompts when eliciting values,
- The effect of iteration on value-based prototypes,
- How methods and techniques can be created for measuring and evaluating value embodiment and value sensitivity (in general and for CAs),
- The impacts of using different value taxonomies,
- How existing value taxonomies can be expanded beyond focusing on the nature of the values themselves and considering other aspects such as user groups, use-cases, CA roles, and value sources or owners,
- How the guidelines provided can generalise to support value embodiment in other types of AI-based systems beyond CAs.

8. Conclusion

This paper introduced the Value-Sensitive Conversational Agent (VSCA) framework and toolkit, aimed at supporting value sensitivity when designing conversational agents (CAs). The framework and toolkit were developed based on systematic literature reviews and interviews with industry practitioners. An evaluation study was conducted using a mixed-methods approach. This consisted of qualitative analysis via design workshops and interviews, which explored the technical utility and value elicitation aspects related to using the VSCA process, and a quantitative survey which examined the value embodiment of the prototypes produced using VSCA by measuring value efficiency, translation, and compatibility. The qualitative analysis focused on whether VSCA increased intentionality and depth in working with values, while the quantitative analysis evaluated VSCA's efficacy in embedding values into prototypes. Fourteen guidelines for CA creators to support practical value embodiment were derived, including prompting value discussions, documenting value-based outcomes, considering LLMs' impacts on values, and identifying value sources/owners. The VSCA framework and toolkit will undergo further updates based on the study results. The overall goal is to provide explicit mechanisms to help practitioners embed diverse, interdisciplinary, and inclusive human values into the CAs they create.

Notes

1. Sadek et al., 2023a, 2023b; Sadek, Calvo, et al., 2024; Sadek, Constantinides, et al., 2024; Sadek, Kallina, et al., 2024.
2. The canvas is accessible at: https://www.conversationdesigninstitute.com/communications/cd-tool-kit.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Leverhulme Trust through the Leverhulme Centre for the Future of Intelligence under Grant RC-2015-067.

ORCID

Malak Sadek http://orcid.org/0000-0001-8284-890X
Rafael A. Calvo http://orcid.org/0000-0003-2238-0684
Céline Mougenot http://orcid.org/0000-0002-3849-163X

Data availability statement

The VSCA Framework and Toolkit are available upon request by contacting the corresponding author.

References

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). *The illusion of artificial inclusion* [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA), Chi '24). Association for Computing Machinery, New York, NY, USA, Article 28612. https://doi.org/10.1145/3613904.3642703

Aguiar-Díaz, I., & Ruiz-Mallorquí, M. V. (2024). Cultural values and internet banking in Europe. The moderating effect of gender. *International Journal of Human–Computer Interaction*. Advance online publication. https://doi.org/10.1080/10447318.2023.2300014

Ahn, J., Campos, F., Nguyen, H., Hays, M., & Morrison, J. (2021). *Co-designing for privacy, transparency, and trust in k-12 learning analytics* [Paper presentation]. Proceedings of the International Learning Analytics and Knowledge Conference (LAK). https://doi.org/10.1145/3448139.3448145

Ali, S. J., Christin, A., Smart, A., & Katila, R. (2023). Walking the walk of ai ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs [Paper presentation]. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL, USA) (FAccT '23). Association for Computing Machinery, New York, NY, USA (pp. 217–226). https://doi.org/10.1145/3593013.3593990

Almutairi, N., Vlahu-Gjorgievska, E., & Win, K. T. (2024). mHealth asthma management app's content creation, stakeholders' values and design features. *International Journal of Human–Computer Interaction*. Advance online publication. https://doi.org/10.1080/10447318.2024.2338326

Alspaugh, S., Zokaei, N., Liu, A., Jin, C., & Hearst, M. (2018). Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 22–31. https://doi.org/10.1109/TVCG.2018.2865040

Alves-Oliveira, P., Lupetti, M., Luria, M., Löffler, D., Gamboa, M., Albaugh, L., Kamino, W., Ostrowski, A., Puljiz, D., Reynolds-Cuéllar, P., Scheunemann, M., Suguitan, M., & Lockton, D. (2021). *Collection of metaphors for human-robot interaction* [Paper presentation]. Proceedings of the Designing Interactive Systems Conference (DIS). https://doi.org/10.1145/3461778.3462060

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). *Guidelines for human-AI interaction* [Paper presentation]. CHI Conference on Human Factors in Computing Systems Proceedings. https://doi.org/10.1145/3290605.3300233

Arrighi, P. A., & Mougenot, C. (2019). Towards user empowerment in product design: A mixed reality tool for interactive virtual prototyping. *Journal of Intelligent Manufacturing*, *30*(2), 743–754. https://doi.org/10.1007/s10845-016-1276-0

Ashfaq, M., Yun, J., & Yu, S. (2021). My smart speaker is cool! Perceived coolness, perceived values, and users' attitude toward smart speakers. *International Journal of Human–Computer Interaction*, *37*(6), 560–573. https://doi.org/10.1080/10447318.2020.1841404

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). Which humans? *PsyArXiv 2023*. https://doi.org/10.31234/osf.io/5b26t

Axelsson, M., Oliveira, R., Racca, M., & Kyrki, V. (2022). Social robot co-design canvases: A participatory design framework. *ACM Transactions on Human-Robot Interaction*, *11*(1), 1–39. Article 3oct 2021. https://doi.org/10.1145/3472225

Balkin, R. S., & Sheperis, C. J. (2011). Evaluating and reporting statistical power in counseling research. *Journal of Counseling & Development*, *89*(3), 268–272. https://doi.org/10.1002/j.1556-6678.2011.tb00088.x

Benkler, N., Mosaphir, D., Friedman, S., Smart, A., & Schmer-Galunder, S. (2023). Assessing LLMs for moral value pluralism. AI Meets Moral Philosophy and Moral Psychology: An Interdisciplinary Dialogue about Computational Ethics', Workshop at NeurIPS 2023. https://doi.org/10.48550/arXiv.2312.10075

Berman, G., Goyal, N., & Madaio, M. (2024). A scoping study of evaluation practices for responsible AI tools: Steps towards effectiveness evaluations [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA,) (CHI '24). Association for Computing Machinery, New

York, NY, USA, Article 294 (p. 2). https://doi.org/10.1145/3613904.3642398

Bingley, W. J., Curtis, C., Lockey, S., Bialkowski, A., Gillespie, N., Haslam, S. A., Ko, R. K. L., Steffens, N., Wiles, J., & Worthy, P. (2023). Where is the human in human-centered AI? Insights from developer priorities and user experiences. *Computers in Human Behavior*, *141*(2023), 107617. https://doi.org/10.1016/j.chb.2022.107617

Borning, A., Friedman, B., Davis, J., & Lin, P. (2005). Informing public deliberation: Value sensitive design of indicators for a large-scale urban simulation. In H. Gellersen, K. Schmidt, M. Beaudouin-Lafon, & W. Mackay (Eds.), *ECSCW 2005* (pp. 449–468). Springer. https://doi.org/10.1007/1-4020-4023-7_23

Boujut, J. F., & Blanco, E. (2003). Intermediary objects as a means to foster co-operation in engineering design. *Computer Supported Cooperative Work (CSCW)*, *12*(2), 205–219. https://doi.org/10.1023/A:1023980212097

Bratteteig, T., & Wagner, I. (2016). Unpacking the notion of participation in participatory design. *Computer Supported Cooperative Work (CSCW)*, *25*(6), 425–475. https://doi.org/10.1007/s10606-016-9259-4

Braun, V., & Clark, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Braun, V., & Clarke, V. (2012). *Thematic analysis* (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

Bubaš, G., Čižmešija, A., & Kovačić, A. (2023). Development of an assessment scale for measurement of usability and user experience characteristics of bing chat conversational AI. *Future Internet*, *16*(1), 4. https://doi.org/10.3390/fi16010004

Byrd, L. (2020). *Tech Inclusion: How culture may enhance conversational AI technology*. https://uxmag.com/articles/tech-inclusion-how-culture-may-enhance-conversational-ai-technology

Cabitza, F., & Simone, C. (2012). Affording mechanisms: An integrated view of coordination and knowledge management. *Computer Supported Cooperative Work (CSCW)*, *21*(2–3), 227–260. https://doi.org/10.1007/s10606-011-9153-z

Calvo, R., & Peters, D. (2017). *Positive computing*. MIT Press.

Candello, H., Pichiliani, M., Pinhanez, C., Vidon, S., & Wessel, M. (2020). *Co-designing a conversational interactive exhibit for children* [Paper presentation]. Proceedings of the ACM Interaction Design and Children Conference: Extended Abstracts. https://doi.org/10.1145/3397617.3397840

Chen, X., Burke, J., Du, R., Hong, M., Jacobs, J., Laban, P., Li, D., Peng, N., Willis, K., Wu, C., & Zhou, B. (2023). Next steps for human-centered generative AI: A technical perspective. *Computing Research Repository (CoRR)*. https://arxiv.org/abs/2306.15774

Chen, Z., Lu, Y., Nieminen, A., & Lucero, M. (2020). Creating a chatbot for and with migrants: Chatbot personality drives co-design activities. In *Proceedings of the ACM Designing Interactive Systems Conference*.

Chow, J. C. L., Sanders, L., & Li, K. (2023). Impact of ChatGPT on medical chatbots as a disruptive technology. *Frontiers in Artificial Intelligence*, *6*(2023), 1166014. https://doi.org/10.3389/frai.2023.1166014

Chow, J. C. L., Wong, V., & Li, K. (2024). Generative pre-trained transformer-empowered healthcare conversations: Current trends, challenges, and future directions in large language model-enabled medical chatbots. *BioMedInformatics*, *4*(1), 837–852. https://doi.org/10.3390/biomedinformatics4010047

Christian, B. (2020). *The alignment problem – Machine learning and human values*. W. W. Norton & Company.

Cockton, G. (2005). *A development framework for value-centred design* [Paper presentation]. CHI Extended Abstracts on Human Factors in Computing Systems. https://doi.org/10.1145/1056808.1056899

Cockton, G. (2006). *Designing worth is worth designing* [Paper presentation]. Proceedings of the Nordic Conference on Human-Computer Interaction. https://doi.org/10.1145/1182475.1182493

Cooper, N., & Zafiroglu, A. (2024). From fitting participation to forging relationships: The art of participatory ML [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 746 (p. 9). https://doi.org/10.1145/3613904.3642775

Curtis, M. J., Bond, R. A., Spina, D., Ahluwalia, A., Alexander, S. P., Giembycz, M. A., Gilchrist, A., Hoyer, D., Insel, P. A., Izzo, A. A., Lawrence, A. J., MacEwan, D. J., Moon, L. D., Wonnacott, S., Weston, A. H., & McGrath, J. C. (2015). Experimental design and analysis and their reporting: new guidance for publication in BJP. *British Journal of Pharmacology*, *172*(14), 3461–3471. https://doi.org/10.1111/bph.12856

Custis, C. (2021). *Operationalizing AI ethics through documentation: ABOUT ML in 2020 and beyond*. https://partnershiponai.org/about-ml-2021/

Czeskis, A., Dermendjieva, I., Yapit, H., Borning, A., Friedman, B., Gill, B., & Kohno, T. (2010). Parenting from the pocket: Value tensions and technical directions for secure and private parent-teen mobile safety. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*.

Dantec, C., Poole, E., & Wyche, S. (2009). Values As lived experience: Evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Darling-Wolf, H., & Patitsas, E. (2024). "Not my priority:" Ethics and the boundaries of computer science identities in undergraduate CS education. In Proceedings of the ACM on Human-Computer Interaction, 8, *CSCW1*, Article 174 (Apr 2024) (p. 28). https://doi.org/10.1145/3641013

Deloitte Digital. (2019). *Conversational AI: The next wave of customer and employee experience*.

Deng, H. W., Guo, B., Devrio, A., Shen, H., Eslami, M., & Holstein, K. (2023). Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice [Paper presentation]. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 377 (p. 18). https://doi.org/10.1145/3544548.3581026

Deng, H. W., Yildirim, N., Chang, M., Eslami, M., Holstein, K., & Madaio, M. (2023). Investigating *practices and opportunities for cross-functional collaboration around* AI *fairness in industry practice* [Paper presentation]. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, (Chicago, IL, USA) (FAccT '23). Association for Computing Machinery, New York, NY, USA (pp. 705–716). https://doi.org/10.1145/3593013.3594037

Denning, P., & Rousse, B. S. (2024). Can machines be in language? *Communications of the ACM*, *67*(3), 32–35. https://doi.org/10.1145/3637629

Denning, T., Borning, A., Friedman, B., Gill, B. T., Kohno, T., & Maisel, W. H. (2010). Patients, pacemakers, and implantable defibrillators: Human values and security for wireless implantable medical devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Dickhaut, D., Janson, A., Söllner, M., & Leimeister, J. M. (2024). Lawfulness by design – Development and evaluation of lawful design patterns to consider legal requirements. *European Journal of Information Systems*, *33*(4), 441–468. https://doi.org/10.1080/0960085X.2023.2174050

DigitalCatapult. (2020). *Lessons in Practical AI Ethics*.

Dindler, C., Krogh, P., Tikær, K., & Nørregård, P. (2022). Engagements and articulations of ethics in design practice. *International Journal of Design*, *16*(2), 47–56. https://doi.org/10.57698/v16i2.04

Dobbe, R., Gilbert, T. K., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, *300*(2021), 103555. https://doi.org/10.1016/j.artint.2021.103555

Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, Z., Hatfield-Dodds, C., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., Ganguli, D. (2024). Towards measuring the representation of subjective global opinions in language models. *Computing Research Repository (CoRR)* (2024). https://arxiv.org/abs/2306.16388

Dutta, S., & Chakraborty, T. (2023). Thus spake ChatGPT. *Communications of the ACM*, *66*(12), 16–19. https://doi.org/10.1145/3616863

Easton, K., Potter, S., Bec, R., Bennion, M., Christensen, H., Grindell, C., Mirheidari, B., Weich, S., de Witte, L., Wolstenholme, D., & Hawley, M. (2019). A virtual agent to support individuals living with physical and mental comorbidities: Co-design and acceptability testing. *Journal of Medical Internet Research*, *21*(5), e12996. https://doi.org/10.2196/12996

Elshan, E., Ebel, P., Söllner, M., & Leimeister, J. M. (2023). Leveraging low code development of smart personal assistants: An integrated design approach with the SPADE method. *Journal of Management Information Systems*, *40*(1), 96–129. https://doi.org/10.1080/07421222.2023.2172776

Elshan, E., Engel, C., Ebel, P., & Siemon, D. (2022). Assessing the reusability of design principles in the realm of conversational agents. In *Proceedings of the International Conference on Design Science Research in Information Systems and Technology (DESRIST)*.

Ezzat, H., Agogué, M., Le Masson, P., Weil, B., & Cassotti, M. (2020). Specificity and abstraction of examples: Opposite effects on fixation for creative ideation. *The Journal of Creative Behavior*, *54*(1), 115–122. https://api.semanticscholar.org/CorpusID:149811990 https://doi.org/10.1002/jocb.349

Faber, I., van Renswouw, L., & Colombo, S. (2024). *A tangible toolkit to uncover clinician's ethical values about AI clinical decision support systems* [Paper presentation]. DRS2024: Boston. https://doi.org/10.21606/drs.2024.862

Fagerlin, A., Pignone, M., Abhyankar, P., Col, N., Feldman-Stewart, D., Gavaruzzi, T., Kryworuchko, J., Levin, C. A., Pieterse, A. H., Reyna, V., Stiggelbout, A., Scherer, L. D., Wills, C., & Witteman, H. O. (2013). Clarifying values: An updated review. *BMC Medical Informatics and Decision Making*, *13* (Suppl 2), S8. https://doi.org/10.1186/1472-6947-13-S2-S8

Faruk, L. I. D., Pal, D., Funilkul, S., Perumal, T., & Mongkolnam, P. (2024). Introducing CASUX: A standardized scale for measuring the user experience of artificial intelligence based conversational agents. *International Journal of Human–Computer Interaction*. Advance online publication. https://doi.org/10.1080/10447318.2024.2359206

Fest, I., Schäfer, M., Dijck, J., & Meijer, A. (2023). Understanding data professionals in the police: A qualitative study of system-level bureaucrats. *Public Management Review*, *25*(9), 1664–1684. https://doi.org/10.1080/14719037.2023.2222734

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying values in technology: Theory and practice. *Information Technology and Moral Philosophy*, 322–346.

Følstad, A., Araujo, T., Law, E. L.-C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., Baez, M., Laban, G., McAllister, P., Ischen, C., Wald, R., Catania, F., Meyer von Wolff, R., Hobert, S., & Luger, E. (2021). Future directions for chatbot research: An interdisciplinary research agenda. *Computing*, *103*(12), 2915–2942. https://doi.org/10.1007/s00607-021-01016-7

Freier, N. (2008). *Children attribute moral standing to a personified agent* [Paper presentation]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/1357054.1357113

Friedman, B., & Hendry, D. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.

Friedman, B., Harbers, M., Hendry, D., van den Hoven, J., Jonker, C., & Logler, N. (2021). Eight grand challenges for value sensitive design from the 2016 Lorentz workshop. *Ethics and Information Technology*, *23*(1), 5–16. https://doi.org/10.1007/s10676-021-09586-y

Friedman, B., Kahn, P., Hagman, J., Severson, R., & Gill, B. (2006). The watcher and the watched: Social judgments about privacy in a public place. *Human-Computer Interaction*, *21*(2), 235–272. https://doi.org/10.1207/s15327051hci2102_3

Fuchsberger, V., Moser, C., & Tscheligi, M. (2012). Values in action (ViA): Combining usability, user experience and user acceptance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35–42. 2021. https://doi.org/10.1016/j.socec.2010.10.008

Garg, R., & Sengupta, S. (2020). Conversational technologies for in-home learning: Using co-design to understand children's and parents' perspectives. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Go FAIR. (2021). *Data based science: FAIR becomes the new normal*. https://www.go-fair.org/2021/01/21/data-based-science-fair-becomes-the-new-normal/

Ha, J., Jeon, H., Han, D., Seo, J., & Oh, C. (2024). CloChat: Understanding how people customize, interact, and experience personas in large language models [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 305 (p. 24). https://doi.org/10.1145/3613904.3642472

Hagendorff, T. (2020). The ethics of AI ethics: an evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hallensleben, S., & Hustedt, C. (2020). *From principles to practice; An interdisciplinary framework to operationalise AI ethics*.

Harbers, M., & Overdiek, A. (2022). Towards a living lab for responsible applied AI. In *Proceedings of the Design Research Society Conference (DRS)*.

Hayes, G. (2020). Inclusive and engaged HCI. *Interactions*, *27*(2), 26–31. https://doi.org/10.1145/3378561

Hedman, J., & Gimpel, G. (2010). The adoption of hyped technologies: a qualitative study. *Information Technology and Management*, *11*(4), 161–175. https://doi.org/10.1007/s10799-010-0075-0

Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine (1982)*, *292*(2022), 114523. https://doi.org/10.1016/j.socscimed.2021.114523

Heo, J., & Lee, U. (2023). Form to flow: Exploring challenges and roles of conversational UX designers in real-world, multi-channel service environments. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–24. https://doi.org/10.1145/3610189

IEEE 7000. (2021). *IEEE standard model process for addressing ethical concerns during system design*. Standard. The Institute of Electrical and Electronics Engineers.

Ikkatai, Y., Hartwig, T., Takanashi, N., & Yokoyama, H. M. (2022). Octagon measurement: Public attitudes toward AI ethics. *International Journal of Human–Computer Interaction*, *38*(17), 1589–1606. https://doi.org/10.1080/10447318.2021.2009669

Jacobs, A., & Wallach, H. (2021). *Measurement and fairness* [Paper presentation]. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT). https://doi.org/10.1145/3442188.3445901

Jakesch, M., Buçinca, Z., Amershi, S., & Olteanu, A. (2022). *How different groups prioritize ethical values for responsible AI* [Paper presentation]. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT). https://doi.org/10.1145/3531146.3533097

Janson, A., Dickhaut, E., & Söllner, M. (2022). *Designing for cultural values: Towards a theory-motivated method for culture-sensitive adaptation of information systems* [Paper presentation]. Proceedings of the 55th Hawaii International Conference on System Sciences. https://hdl.handle.net/10125/79904 https://doi.org/10.24251/HICSS.2022.567

Jansson, D. G., & Smith, S. M. (1991). Design fixation. *Design Studies*, *12*(1), 3–11. https://doi.org/10.1016/0142-694X(91)90003-F

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Jonas, W. (2018). *A cybernetic model of design research: Towards a trans-domain of knowing*. Routledge.

Kayacik, C., Chen, S., Noerly, S., & Holbrook, J. (2019). Identifying the intersections: User experience + research scientist collaboration in a

generative machine learning interface. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Keys, Y., Silverman, S., & Evans, J. (2017). Identification of tools and techniques to enhance interdisciplinary collaboration during design and construction projects. *Herd*, *10*(5), 28–38. https://doi.org/10.1177/1937586716684135

Kim, J., & Kim, M. (2024). Influences of ethical values of e-travel agents on authenticity and trust formations: The moderating role of privacy concerns. *International Journal of Human–Computer Interaction*, *40*(20), 6442–6457. https://doi.org/10.1080/10447318.2023.2254638

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*(2024), 100225. https://doi.org/10.1016/j.caeai.2024.100225

Knowles, B., Richards, J., & Weiss, G. (2023). Trusted AI: Last in a series on systems and trust. ACM technology policy tech briefs, 9. https://doi.org/10.1145/3641524

Koesten, L., Kacprzak, E., Tennison, J., & Simperl, E. (2019). Collaborative practices with structured data: do tools support what users need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Kroes, P., & Van de Poel, I. (2020). *Design for values and the definition, specification, and operationalization of values*. Springer Reference.

Kross, S (2022). *Prototyping the developer experience for data science practitioners and instructors* [Ph.D. dissertation]. University of California.

Kwon, H., Choi, Y. J., Lee, S., & Lee, S. (2024). Unveiling the inherent needs: GPT builder as participatory design tool for exploring needs and expectation of ai with middle-aged users [Paper presentation]. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 358 (p. 6). https://doi.org/10.1145/3613905.3650794

Lee, C. (2020). Boundary objects and coordination for creativity and innovation. Presentation at the PAIR Symposium and can be accessed at: https://www.youtube.com/watch?v=zpzMlIjKSbc&t=3839s.

Lee, S., Lee, J., & Lee, K. (2017). *Designing intelligent assistant through user participations* [Paper presentation]. Proceedings of the Conference on Designing Interactive Systems. https://doi.org/10.1145/3064663.3064733

Lewandowski, T., Delling, J., Grotherr, C., & Böhmann, T. (2021). State-of-the-art analysis of adopting AI-based conversational agents in organizations: A systematic literature review. In Proceedings of Pacific Asia Conference on Information Systems (PACIS).

Lewandowski, T., Heuer, M., Vogel, P., & Böhmann, T. (2022). Design knowledge for the lifecycle management of conversational agents. *Proceedings of Wirtschaftsinformatik (WI)*.

Li, B., Wang, A., Strachan, P., Séguin, J. A., Lachgar, S., Schroeder, K. C., Fleck, M. S., Wong, R., Karthikesalingam, A., Natarajan, V., Matias, Y., Corrado, G. S., Webster, D., Liu, Y., Hammel, N., Sayres, R., Semturs, C., & Schaekermann, M. (2024). *Conversational AI in health: Design considerations from a Wizard-of-Oz dermatology case study with users, clinicians and a medical LLM* [Paper presentation]. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 8810. https://doi.org/10.1145/3613905.3651891

Lim, Y., & Kim, B. (2022). Value-drive design approach to envision speculative futures. In *Proceedings of the Design Research Society Conference (DRS)*.

Lopatovska, I., Turpin, O., Davis, J., Connell, E., Denney, C., Fournier, H., Ravi, A., Yoon, J. H., & Parasnis, E. (2022). *Capturing teens' voice in designing supportive agents* [Paper presentation]. Proceedings of the 4th Conference on Conversational User Interfaces. https://doi.org/10.1145/3543829.3543838

Luger, E., & Sellen, A. (2016). "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Luria, M. (2018). *Designing robot personality based on fictional sidekick characters* [Paper presentation]. Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction. https://doi.org/10.1145/3173386.3176912

Luria, M., Oden Choi, J., Karp, R. G., Zimmerman, J., & Forlizzi, J. (2020). *Robotic futures: Learning about personally-owned agents through performance* [Paper presentation]. Proceedings of the ACM Designing Interactive Systems Conference (DIS). https://doi.org/10.1145/3357236.3395488

Manders-Huits, N. (2011). What values in design? The challenge of incorporating moral values into design. *Science and Engineering Ethics*, *17*(2), 271–287. https://doi.org/10.1007/s11948-010-9198-2

Manders-Huits, N., & Zimmer, M. (2009). Values and pragmatic action: The challenges of introducing ethical intelligence in technical design communities. *The International Review of Information Ethics*, *10*(2009), 37–44. https://doi.org/10.29173/irie87

Martin, D. (2020). Causal maps as boundary objects for understanding AI development "territories". *Presentation at the PAIR Symposium and Can Be Accessed at*: https://www.youtube.com/watch?v=J2smNXwflYs&t=3793s.

Masson, D., Malacria, S., Casiez, G., & Vogel, D. (2024). *DirectGPT: A direct manipulation interface to interact with large language models* [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 975 (p. 16). https://doi.org/10.1145/3613904.3642462

McGrath, Q. P. (2024). Unveiling the ethical positions of conversational AIs: A study on OpenAI's ChatGPT and Google's Bard. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00433-6

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). *Does ACM's code of ethics change ethical decision making in software development?* [Paper presentation]. Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (Lake Buena Vista, FL, USA) (ESEC/FSE 2018). Association for Computing Machinery, New York, NY, USA (pp. 729–733). https://doi.org/10.1145/3236024.3264833

Mepham, B., Kaiser, M., Thortensen, E., Tomkins, S., & Millar, K. (2006). Ethical matrix manual. *Agricultural and Forest Meteorology*.

Miller, J., Friedman, B., Jancke, G., & Gill, B. (2007). *Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system* [Paper presentation]. Proceedings of the 2007 International ACM Conference on Supporting Group Work. https://doi.org/10.1145/1316624.1316668

Mitchell, M. (2020). Intentional ignorance is a value-laden choice. *Presentation at the PAIR Symposium*.

Monteiro, M., & Salgado, L. (2023). Conversational Agents: a survey on culturally informed design practices. *Journal on Interactive Systems*, *14*(1), 33–46. 2023. https://doi.org/10.5753/jis.2023.2961

Moore, R. (2018). *A natural conversation framework for conversational UX design* (pp. 181–204). Springer International Publishing.

Moore, R., Hosn, R., & Arora, A. (2016). *The machinery of natural conversation and the design of conversational machines* [Paper presentation]. American Sociological Association Annual Meeting.

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: barriers, enablers and next steps. *AI & Society*, *38*(1), 411–423. https://doi.org/10.1007/s00146-021-01308-8

Murad, C., Candello, H., & Munteanu, C. (2023). *What's the talk on VUI guidelines? A meta-analysis of guidelines for voice user interface design* [Paper presentation]. Proceedings of the ACM Conference on Conversational User Interfaces (CUI). https://doi.org/10.1145/3571884.3597129

Palmer, A., & Schwan, D. (2024). More process, less principles: The ethics of deploying AI and robotics in medicine. *Cambridge Quarterly of Healthcare Ethics*, *33*(1), 121–134. https://doi.org/10.1017/S0963180123000087

Park, J. S., Karahalios, K., Salehi, N., & Eslami, M. (2022). Power dynamics and value conflicts in designing and maintaining socio-technical algorithmic processes. *Proceedings of the ACM on Human-*

Computer Interaction, 6(CSCW1), 110. https://doi.org/10.1145/3512957

Pennington, D. (2010). The dynamics of material artifacts in collaborative research teams. *Computer Supported Cooperative Work (CSCW), 19*(2), 175–199. https://doi.org/10.1007/s10606-010-9108-9

Piorkowski, D., Park, S., Wang, A., Wang, D., Muller, M., & Portnoy, F. (2021). How AI developers overcome communication challenges in a multidisciplinary team: A case study. (2021). *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–25. https://doi.org/10.1145/3449205

Pistilli, G., Leidinger, A., Jernite1, Y., Kasirzadeh, A., Luccioni, A. S., & Mitchell, M. (2024). CIVICS: Building a dataset for examining culturally-informed values in large language models. *Computing Research Repository (CoRR)*. https://arxiv.org/abs/2405.13974

Pommeranz, A., Detweiler, C., Wiggers, P., & Jonker, C. (2012). Elicitation of situated values: Need for tools to help stakeholders and designers to reflect and communicate. *Ethics and Information Technology*, *14*(4), 285–303. https://doi.org/10.1007/s10676-011-9282-6

Preston, S. D., & de Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *The Behavioral and Brain Sciences*, *25*(1), 1–20. https://doi.org/10.1017/s0140525x02000018

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

RASA. (2021). Conversational AI in customer experience.

Redaelli, I., & Carassa, A. (2018). New perspectives on plans: Studying planning as an instance of instructed action. *Computer Supported Cooperative Work (CSCW)*, *27*(1), 107–148. https://doi.org/10.1007/s10606-017-9278-9

Reinecke, K., & Bernstein, A. (2013). Knowing what a user likes: A design science approach to interfaces that automatically adapt to culture. *MIS Quarterly*, *37*(2), 427–453. http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=87371385&site=ehost-live https://doi.org/10.25300/MISQ/2013/37.2.06

Rong, A., & Li, G. (2023). Integrating artificial intelligence in design: An exploration of ChatGPT's role in inclusive design research. In *Proceedings of the 2023 5th International Conference on Literature, Art and Human Development (ICLAHD 2023)* (pp. 910–924). Atlantis Press. https://doi.org/10.2991/978-2-38476-170-8_101

Russell, S. (2019). *Human compatible: AI and the problem of control*. Allen Lane.

Sadek, M., Calvo, R. A., & Mougenot, C. (2023a). Co-designing conversational agents: A comprehensive review and recommendations for best practices. *Design Studies*, *89*(2023), 101230. https://doi.org/10.1016/j.destud.2023.101230

Sadek, M., Calvo, R. A., & Mougenot, C. (2023b). *Trends, challenges and processes in conversational agent design: Exploring practitioners' views through semi-structured interviews* [Paper presentation]. Proceedings of the 5th International Conference on Conversational User Interfaces. https://doi.org/10.1145/3571884.3597143

Sadek, M., Calvo, R. A., & Mougenot, C. (2024). Designing value-sensitive AI: A critical review and recommendations for socio-technical design processes. *AI Ethics*, *4*, 949–967. https://doi.org/10.1007/s43681-023-00373-7

Sadek, M., Constantinides, M., Quercia, D., & Mougenot, C. (2024). *Guidelines for integrating value sensitive design in responsible AI toolkits* [Paper presentation]. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3613904.3642810

Sadek, M., Kallina, E., Bohné, T., Mougenot, C., Calvo, R. A., & Cave, S. (2024). Challenges of responsible AI in practice: scoping review and recommended actions. *AI & Society*. https://doi.org/10.1007/s00146-024-01880-9

Sanders, E., & Stappers, P. (2008). Co-creation and the new landscapes of design. *CoDesign*, *4*(1), 5–18. https://doi.org/10.1080/15710880701875068

Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., Newnham, G., Hajkowicz, S., Robinson, C., & Hansen, D. (2023). AI ethics principles in practice: Perspectives of designers and developers. *IEEE Transactions on Technology and Society*, *4*(2), 171–187. https://doi.org/10.1109/TTS.2023.3257303

Schmitt, A., Zierau, N., Janson, J. M., & Leimeister, A. (2023). The role of AI-based artifacts' voice capabilities for agency attribution. *Journal of the Association for Information Systems*, *24*(4), 980–1004. https://doi.org/10.17705/1jais.00827

Schöbel, S., Schmitt, A., Benner, D., Saqr, M., Janson, A., & Leimeister, J. M. (2024). Charting the evolution and future of conversational agents: a research agenda along five waves and new frontiers. *Information Systems Frontiers*, *26*(2), 729–754. https://doi.org/10.1007/s10796-023-10375-9

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O., & Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, *103*(4), 663–688. https://doi.org/10.1037/a0029393

Shilton, K. (2013). Values levers: Building ethics into design. *Science, Technology, & Human Values*, *38*(3), 374–397. https://doi.org/10.1177/0162243912436985

Showkat, D., & Baumer, E. (2022). *"It's like the value system in the loop": Domain experts' values expectations for NLP automation* [Paper presentation]. Proceedings of the Designing Interactive Systems (DIS) Conference. https://doi.org/10.1145/3532106.3533483

Siddique, S., & Chow, J. C. L. (2021). Machine learning in healthcare communication. *Encyclopedia*, *1*(1), 220–239. https://doi.org/10.3390/encyclopedia1010021

Sio, U. N., Kotovsky, K., & Cagan, J. (2015). Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies*, *39*(2015), 70–99. https://doi.org/10.1016/j.destud.2015.04.004

Sleeswijk, V. (2018). Structuring Roles in Research through Design Collaboration. In *Proceedings of the Design Research Society, DRS 2018: Design as a Catalyst for Change*.

Smith, F., Sadek, M., & Mougenot, C. (2023). *Empowering end-users in co-designing AI: an AI literacy card-based toolkit for non-technical audiences* [Paper presentation]. Proceedings of the 36th International BCS Human-Computer Interaction Conference. https://doi.org/10.14236/ewic/BCSHCI2023.2

Sonntag, M., Schuir, J., Mehmann, J., & Teuteberg, F. (2023). Customer service ethics in the metaverse: A value-sensitive design science approach. *International Journal of Human–Computer Interaction*. Advance online publication. https://doi.org/10.1080/10447318.2023.2298533

Steen, M. (2013). Co-design as a process of joint inquiry and imagination. *Design Issues*, *29*(2), 16–28. https://doi.org/10.1162/DESI_a_00207

Stray, J., Vendrov, I., Nixon, J., Adler, S., & Hadfield-Menell, D. (2020). What are you optimizing for? Aligning Recommender Systems with Human Values. In *Proceedings of the ICML 2020 Participatory Approaches to Machine Learning Workshop*.

Strikwerda, L., van Steenbergen, M., van Gorp, A., Timmers, C., & van Grondelle, J. (2022). The value sensitive design of a preventive health check app. *Ethics and Information Technology*, *24*(3), 38. https://doi.org/10.1007/s10676-022-09662-x

Tanner, B., & Feltz, A. (2022). Comparing effects of default nudges and informing on recycled water decisions. *Journal of Experimental Psychology. Applied*, *28*(2), 399–411. https://doi.org/10.1037/xap0000406

Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, J. M., & Bernstein, A. (2021). *Female by default? – Exploring the effect of voice assistant gender and pitch on trait and trust attribution* [Paper presentation]. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3411763.3451623

Toloka. (2023). *The history, timeline, and future of LLMs*. https://toloka.ai/blog/history-of-llms/

Umbrello, S. (2018). The moral psychology of value sensitive design: The methodological issues of moral intuitions for responsible innovation. *Journal of Responsible Innovation*, *5*(2), 186–200. https://doi.org/10.1080/23299460.2018.1457401

Umbrello, S. (2021). Conceptualizing policy in value sensitive design: A machine ethics approach. In S. J. Thompson (Ed.), *Machine law,*

ethics, and morality in the age of artificial intelligence (pp. 108–125). IGI Global. https://doi.org/10.4018/978-1-7998-4894-3.ch007

Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, *1*(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3

van de Poel, I. (2013). *Translating values into design requirements*. Springer.

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

Varanasi, R. A., & Goyal, N. (2023). *"It is currently hodgepodge": Examining AI/ML practitioners' challenges during co-production of responsible AI values* [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3544548.3580903

Venkatesh, V. (2022). Adoption and use of AI tools: A research agenda grounded in UTAUT. *Annals of Operations Research*, *308*(1-2), 641–652. https://doi.org/10.1007/s10479-020-03918-9

Vera, Q., & Muller, M. (2019). Enabling value sensitive AI systems through participatory design fictions. *CoRR*. https://doi.org/10.48550/arXiv.1912.07381

Verdiesen, I., & Dignum, V. (2022). Value elicitation on a scenario of autonomous weapon system deployment: A qualitative study based on the value deliberation process. *AI and Ethics*, 1–14. https://doi.org/10.1007/s43681-022-00211-2

Vernim, S., Bauer, H., Rauch, E., Ziegler, M., & Umbrello, S. (2022). *A value sensitive design approach for designing AI-based worker assistance systems in manufacturing* [Paper presentation]. International Conference on Industry 4.0 and Smart Manufacturing.

Videla, A. (2023). Echoes of intelligence: Textual interpretation and large language models. *Communications of the ACM*, *66*(11), 38–43. https://doi.org/10.1145/3610283

Vines, J., Clarke, R., Wright, P., McCarthy, J., & Olivier, P. (2013). Configuring participation: On how we involve people in design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Visser, F. S., Stappers, P. J., van der Lugt, R., & Sanders, E. B. N. (2005). Contextmapping: Experiences from practice. *CoDesign*, *1*(2), 119–149. https://doi.org/10.1080/15710880500135987

Vogelsang, A., & Borg, M. (2019). Requirements engineering for machine learning: Perspectives from data scientists. In *Proceedings of the IEEE International Requirements Engineering Conference*.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review*, *2*(2019).

Wahde, M., & Virgolin, M. (2022). *Conversational agents: Theory and applications*. World Scientific Publishing Company.

Wambsganss, T., Höch, A., Zierau, N., & Söllner, M. (2021). Ethical design of conversational agents: Towards principles for a value-sensitive design. In Lecture notes in information systems and organisation.

Wang, L., Chen, C., & Huh-Yoo, J. (2023). Investigating the synonyms of conversational agents to aid cross-disciplinary ca research [Paper presentation]. Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI EA '23). Association for Computing Machinery, New York, NY, USA, Article 190 (p. 10). https://doi.org/10.1145/3544549.3585640

Wang, Z. J., Kulkarni, C., Wilcox, L., Terry, M., & Madaio, M. (2024). Farsight: Fostering responsible AI awareness during AI application prototyping [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 976 (p. 40). https://doi.org/10.1145/3613904.3642335

Weber, P., Krings, K., Nießner, J., Brodesser, S., & Ludwig, T. (2021). *FoodChattAR: Exploring the design space of edible virtual agents for human-food interaction* [Paper presentation]. Proceedings of the ACM Designing Interactive Systems Conference (DIS). https://doi.org/10.1145/3461778.3461998

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). Sociotechnical safety evaluation of generative AI systems. *arXiv:2310.11986*, [cs.AI]. https://doi.org/10.48550/arXiv.2310.11986

West, S., Whittaker, M., & Crawford, K. (2019). *Discriminating systems: Gender, race, and power*.

Wester, J., Schrills, T., Pohl, H., & van Berkel, N. (2024). "As an AI language model, I cannot": Investigating LLM Denials of User Requests [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24),. Association for Computing Machinery, New York, NY, USA, Article 979 (p. 14). https://doi.org/10.1145/3613904.3642135

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Wilkes, S., Wongsriruksa, S., Howes, P., Gamester, R., Witchel, H., Conreen, M., Laughlin, Z., & Miodownik, M. (2016). Design tools for interdisciplinary translation of material experiences. *Materials & Design*, *90*(2016), 1228–1237. https://doi.org/10.1016/j.matdes.2015.04.013

Winecoff, A., & Watkins, E. (2022). *Artificial concepts of artificial intelligence: Institutional compliance and resistance in AI startups* [Paper presentation]. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3514094.3534138

Wolff, R., Hobert, S., & Schumann, M. (2022). *Chatbot introduction and operation in enterprises – A design science research-based structured procedure model for chatbot projects* [Paper presentation]. Proceedings of the Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2022.721

Wong, R., Madaio, M., & Merrill, N. (2022). Seeing Like a toolkit: How toolkits envision the work of AI ethics. *Computing Research Repository (CoRR)*, https://arxiv.org/abs/2202.08792

Xu, L., Sanders, L., Li, K., & Chow, J. C. L. (2021). Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review. *JMIR Cancer*, *7*(4), e27850. 2021 https://doi.org/10.2196/27850

Yang, Q., Scuito, A., Zimmerman, J., Forlizzi, J., & Steinfeld, A. (2018). *Investigating how experienced UX designers effectively work with machine learning* [Paper presentation]. Proceedings of the CHI Conference on Designing Interactive Systems (DIS). https://doi.org/10.1145/3196709.3196730

Yi, X., Yao, J., Wang, X., Xie, X. (2023). Unpacking the ethical value alignment in big models. *Computing Research Repository (CoRR)*. https://arxiv.org/abs/2310.17551

Yildirim, N., Kass, A., Tung, T., Upton, C., Costello, D., Giusti, R., Lacin, S., Lovic, S., O'Neill, J., Meehan, R., Ó Loideáin, E., Pini, A., Corcoran, M., Hayes, J., Cahalane, D., Shivhare, G., Castoro, L., Caruso, G., Oh, C., … Zimmerman, J. (2022). *How experienced designers of enterprise applications engage AI as a design material* [Paper presentation]. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3491102.351749

Yildirim, N., Pushkarna, M., Goyal, N., Wattenberg, M., & Viégas, F. (2023). *Investigating how practitioners use human-AI guidelines: A case study on the people + AI guidebook*. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3580900

Yokoi, R., & Nakayachi, K. (2021). The effect of value similarity on trust in the automation systems: A case of transportation and medical care. *International Journal of Human–Computer Interaction*, *37*(13), 1269–1282. https://doi.org/10.1080/10447318.2021.1876360

Yoo, D. W., Woo, H., Pendse, S. R., Lu, N. Y., Birnbaum, M. L., Abowd, G. D., & De Choudhury, M. (2024). Missed opportunities for human-centered AI research: Understanding stakeholder collaboration in mental health AI research. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–24. Article 95 (apr 2024). https://doi.org/10.1145/3637372

Yoo, D., Huldtgren, A., Woelfer, J., Hendry, D., & Friedman, B. (2013). A value sensitive action-reflection model: Evolving a co-design space

with stakeholder and designer prompts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*.
Yurrita, M., Murray-Rust, D., Balayn, A., & Bozzon, A. (2022). *Towards a multi-stakeholder value-based assessment framework for algorithmic systems* [Paper presentation]. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT). https://doi.org/10.1145/3531146.3533118
Zdanowska, S., & Taylor, A. (2022). A study of UX practitioners roles in designing real-world, enterprise ML systems [Paper presentation]. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 531 (p. 15). https://doi.org/10.1145/3491102.3517607
Zhang, A., Muller, M., & Wang, D. (2020). How do data science workers collaborate? *Proceedings of the ACM on Human-Computer Interaction*, *4* (CSCW1), 1–23. https://doi.org/10.1145/3392826
Zhu, H., Yu, B., Halfaker, A., & Terveen, L. (2018). Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–23. https://doi.org/10.1145/3274463

About the authors

Malak Sadek is currently pursuing a PhD at the Dyson School of Design Engineering at Imperial College London. Her research focuses on collaborative value-sensitive conversational AI design. Malak holds a MSc in Human-Computer Interaction from St Andrews University, and a BSc in Computer Engineering from The American University in Cairo.

Rafael A. Calvo is a professor and director of research at the Dyson School of Design Engineering at Imperial College London. Rafael is also co-lead at the Leverhulme Centre for the Future of Intelligence, and chief investigator at the Australian Research Hub on Digital Enhanced Living.

Céline Mougenot is a senior lecturer (associate professor) in collaborative design at the Dyson School of Design Engineering at Imperial College London, Céline is also director of the Collective Innovation Laboratory, co-lead of the Women's Health Network and the head of programme for the MSc in Design with Behavioural Science.

Appendix A. Interview study questions

These questions acted as a guide/prompt for the interviewer and were not necessarily asked in the order listed below. Several bullet points also contain multiple questions around the same topic; in the interview these are broken down into several questions, in order not to confuse interviewees.

A1. Questions for CA creators

A1.1. Technical utility

- What do you think is the most significant/memorable moment you recall and why?
- Please describe how you incorporated the toolkit in the workshop across the three activities. Did you use any specific strategies?
- What was the worst thing about the workshop or the thing that just did not work in your opinion?
- When and how would you see yourself using toolkit during your current workflows/tasks/studies?
- How would you use the toolkit alongside other tools you currently use? Would it replace any of them?
- Do you think you would re-use the outcomes several times throughout the design process? Would they help iteration? Does it cover all major phases? How/Why?
- Would you need to adapt or change any of your current practices to incorporate the use of this toolkit?
- Do you think the artefacts act as documentation? Can you see your team using them as such?

A1.2. Perceptions and attitudes

- How do you feel the workshop affected you in terms of sensitivity to others' values?
- What is one thing you have learnt?
- How do you feel about the level of guidance? Was it practical and straightforward to follow? Was there enough guidance? Could you use this unfacilitated?

A1.3. Challenges

- Did toolkit help you overcome any specific challenges?
- Were there any challenges where the toolkit did not help, or any support you needed that they did not provide?
- Discuss the usefulness and issues with use-cases, design requirements, conversation flows, roleplay dialogues and discussions.

A1.4. Suggestions

- Do you have suggestions or areas for improvement with the toolkit?

A2. Questions for CA users

A2.1. Value elicitation

- What do you think is the most significant/memorable moment you recall and why?
- How well do you feel like your values were integrated along the process of the workshop, and in the final script (ask for specific examples / moments in the workshop)?
- How do you feel the different outcomes (sample roleplay dialogue, design requirements, use-cases) respected your values? Please elaborate.
- Can you provide an example of a use-case and explain how it evolved and respected your value?
- Do you feel that the team will be able to respect your values moving forward?

A2.2. Perceptions and attitudes

- How do you feel the toolkit influenced your perception of values and putting them in tech?
- How do you feel the toolkit influenced your values?
- What was the worst thing about the workshop or the thing that just did not work in your opinion?

A2.3. Collaboration and participation

- Did you feel that you were informed enough to be able to participate meaningfully and as much as you'd have liked? Were you able to understand everything well?
- Did you feel that you were heard? That you were able to express your ideas?
- How did you feel about the artefacts in terms of helping to build common ground or being on the same page as the CA creators?
- What would you do differently or what do you find problematic?

A2.4. Suggestions

- Do you have suggestions or areas for improvement with the toolkit?