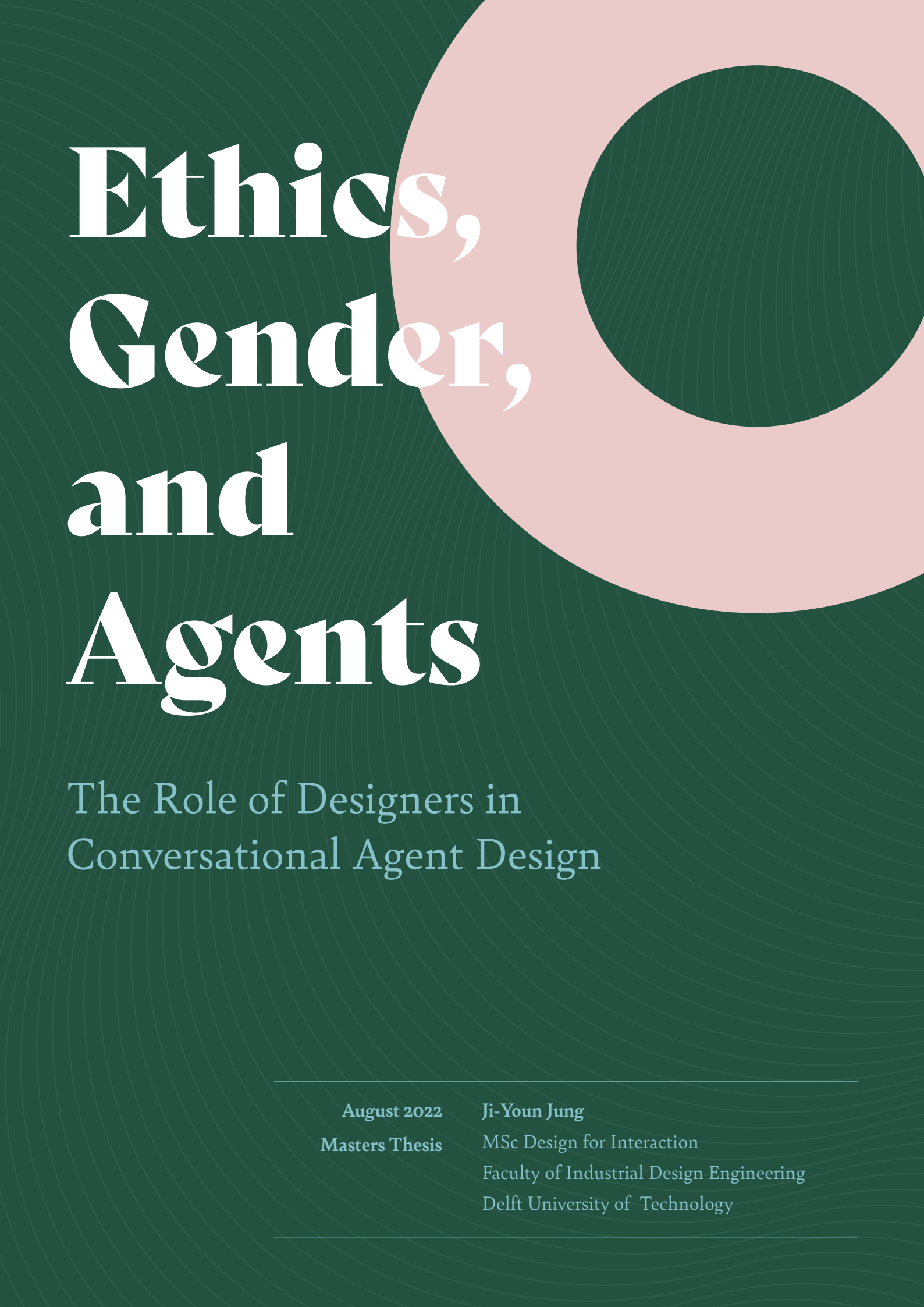


The background features a dark green field with a pattern of fine, concentric white circles. Overlaid on this is a large, light pink circle on the right side, which contains a smaller, solid dark green circle.

Ethics, Gender, and Agents

The Role of Designers in
Conversational Agent Design

August 2022
Masters Thesis

Ji-Youn Jung
MSc Design for Interaction
Faculty of Industrial Design Engineering
Delft University of Technology

Author

Ji-Youn Jung
j.y.jung@student.tudelft.nl
jiyounjung.com

Master Thesis

August 2022
MSc. Design for Interaction
Faculty of Industrial Design Engineering
Delft University of Technology

Graduation Committee**Chair:**

Prof.dr.ir. Alessandro Bozzon
Knowledge and Intelligence Design,
Department of Sustainable Design Engineering (SDE)

Mentor 1:

Dr. Dave Murray-Rust
Associate Professor, Human-Algorithm Interaction
Department of Human-Centered Design (HCD)

Mentor 2:

Dr. Ujwal Gadiraju
Assistant Professor, Web Informatics Systems (WIS)
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS/EWI)

Ethics, Gender, and Metaphor:**The Role of Designers in
Conversational Agent Design**

Preface

Growing up in the late 90s' and early 2000s', most of my childhood nostalgia is strongly attached to cyberspace. Such a virtual environment taught me things I did not know, showed me things I've never felt before, and enabled me to create things I wouldn't have imagined that I could make.

Then when I was studying design and cognitive science for my bachelor's, one peculiar event I came across during listening to one of the student extra-curricular club presentation mesmerized me and led me to pursue Human-Computer Interaction and to investigate this thesis topic. The event was a funeral held on one of the altars at Kofukuji Temple in Japan. But unlike any usual funeral, it was a funeral for the 62 AIBO robot dogs (figure 1) [1].

This event fascinated me for two reasons. First, what made these people get so attached to this "toy" machine that only had very limited performance? Second, how do people apply this robot dog, which was far from a living creature in the sense that it does not breathe or eat, to a funeral ceremony where people remember and respect "the dead"?

I started to dig deep into this phenomena, and I learned a theory of Computers As Social Actors (CASA) theory [2-3], where a number of researchers proved how humans treat interactions with computers, televisions, and new media as social and natural interactions with other human beings. Moreover, they seem to unconsciously anthropomorphize such technology that is basically an assembly of cold metal circuits.

Then, what would happen when such metal objects start to understand human language and react and interact with people in their natural language, just as

Alan Turing imagined [4]? What kind of changes will it bring, and what will be the consequences of such changes?

To make such a conversational user interface (CUI) truly meaningful, designers have a fair share of duty to shape the interaction to be engaging, enabling, and ethical.

This thesis documented my journey trying to figure out and fulfill a small piece in the process.



Figure 1. A funeral for 'Aibo' robot dogs at a temple near Tokyo [1].

Acknowledgement

I would like to show my gratitude to everyone who has given me guidance, support, and delight along my journey of pursuing a Master's degree.

First of all, I would like to express my highest gratitude to my supervisory team, Dr.Ujwal Gadiraju, Dr.Dave Murray-Rust, and Dr.Alessandro Bozzon. Thank you, Ujwal, for your extremely helpful advice and for providing me with many opportunities. Thank you, Dave, for your inspiring ideas and warming support. Thank you, Alessandro, for your amazing feedback that helped me sharpen my scientific viewpoint and showed me a land of opportunities.

Thank you to everyone who has participated in my research, including crowd-workers and some of my fellow students in the Industrial Design Engineering faculty who landed eyes to proofread the experiment materials.

I would like to thank StudioLab for having me during the graduation period, giving me a great working environment, and welcoming me into a fantastic community that they have created. I am particularly grateful to Aadjan van der Helm and Ianus Keller for your inspiring leadership.

I would also like to thank the amazing and inspiring people who provided me mentorship during my studies. Special thanks to Dr. Sihang Qiu for your kind and amazing mentorship, Dr. Yen-Chia Hsu for always being willing to give me advice about the HCI field, and Dr. Valentijn Visch for your mentorship during my Honours Project that lasted one and half years during my studies.

I truly appreciate spending time with my friends in the Netherlands. Special thanks to Kay, Jackie, Ahni, Yeon-ju, Yeun, Sueyoon, Elizabeth, and Francis.

Last but not least, I am forever grateful to my family for their endless love and support. Thank you for always having faith in me and giving me so many opportunities in life.

Ji-Youn Jung
August, 2022
Delft, the Netherlands.

Summary

Research has shown how people anthropomorphize conversational agents (CA) and unconsciously bring their gender stereotypes into human-agent interaction. For this reason, there has been a long lasted dilemma on whether designers should design CAs that conform to or violate stereotypical expectations. Despite the urgency and importance of navigating through this dilemma, how to better design the gender identity of CA is still an open research question. In this thesis, we describe the problem space of CA identity design and argue that we can calibrate the gender effect by manipulating a metaphor we attach to the CAs. To this end, we approached these research questions from three angles: (1) developing a framework to address the ethical dilemma in CA identity design, (2) evaluating the effect of gender and metaphor in chatbot profiles, and (3) calibrating gender stereotyping through metaphor manipulation.

We reviewed previous literature on agent gender design and identified the research gaps. Afterward, we analyzed the CA gender design dilemma that CA designers experience in three layers. Finally, we propose dialogical ethics as a potential ethical framework to help designers navigate and articulate their design practice.

We investigate how people perceive chatbot profiles with different gender markers and a metaphor in the context of a conversational recommender system. To facilitate our research, we conducted a mixed-method study where we collected users' quantitative and qualitative answers. Our study reveals how textual metaphor still persisted when it was attached to visual and gendered cues. Moreover, we show how metaphor showed statistical significance between all conditions, while the gender of CA only differed in perceived warmth.

While CAs of the same gender can manifest endless nuances by adopting different metaphors, to the best of our knowledge, no research has been reported to investigate the effect of CA gender when they manifest different metaphors. To this end, we conducted a qualitative user study in human decision-making.

With our work, we contribute novel knowledge in ethical CA identity design. This thesis concludes with a discussion of the findings of our work and a few directions for further research.



Table of Contents

Preface	2
Acknowledgement	3
Summary	4
Table of Contents	6
Chapter 1. Introduction	8
1.1 Problem Context	10
1.2 Thesis Outline	13
1.3 Research Questions & Contributions	14
1.4 Research Methodology	17
1.5 Project Overview	19
Chapter 2. Background	20
2.1 Conversational Agent (CA) Design	22
2.2 CAs and Gender	29
2.3 Ethical Frameworks	31
2.4 Discussion	33
Chapter 3. Ethical Identity Design Framework	34
3.1 Introduction	36
3.2 Analyzing Ethical CA Design Dilemma	38
3.3 Utilitarian vs. Dialogical Ethics	40
3.4 CUI@CHI Workshop	42
3.5 Discussion	43
3.6 Conclusion	45

Chapter 4. Crafting the agents	46
4.1 Introduction	48
4.2 Methods	49
4.3 Survey Result	54
4.4 Thematic Analysis Result	57
4.5. Discussion	61
4.6 Summary	63
Chapter 5. Can We Overcome the Stereotype?	64
5.1 Introduction	66
5.2 Related Work	68
5.3 Method	70
Chapter 6. Thesis Discussion & Future Work	74
6.1. Summary of Contributions	75
6.2 Design Guidelines	76
6.3 Future Work	79
References	80
Appendices	90

Chapter 1.

Introduction

This chapter introduces key context of the thesis and explain the scope by defining the problem at hand. Moreover, it describes research questions and research approach to address the questions. It concludes by picturing the context of this projcet with invovled stakeholders.

1.1 Problem Context

1.2 Thesis Outline

1.3 Research Questions & Contributions

- Developing Framework to Address Ethical Dilemma in CA Identity Design
- Evaluating the Effect of Gender and Metaphor in Chatbot Profile
- Calibrate Gender Stereotyping through Metaphor Manipulation

1.4 Research Methodology

- Integrative review method
- Mixed-method study
- Quantitative empirical user study
- Crowdsourcing

1.5 Project Overview

1.1 Problem Context

Unlike traditional text-based user interface or graphical user interface (GUI) that requires a learning curve for the user, conversational user interface (CUI) has been endowed as the next natural form of human-computer interaction (HCI) as it allows users to communicate in their natural language. On this account, CUI is becoming ever-more common in everyday lives, in the form of personal assistants (e.g., Apple's Siri, Microsoft's Cortana, Amazon's Alexa, Google's Google Assistant), or the form text-based chatbots (e.g., Microsoft's Xiaoice, HelloFresh's Freddy, KLM's Service bot).

On a similar note, making a recommender system conversational has also been suggested as overcoming a few limitations of the GUI recommender interface [22-24]. As a set of recommendations can be highly context-dependent, it is hard for systems to determine users' current needs solely on users' past interactions. However, by making such recommender conversational, conversational agents (CAs) can conduct a multi-turn dialogue with users, and the system can elicit the detailed preferences of users.

While the conversational agents (CAs) that facilitate natural language conversations with human users are strictly a computer, people anthropomorphize these agents and treat these conversational agents as social actors [109] (CASA paradigm) [3]. Copious research indicates how anthropomorphizing CAs affects users' interactions, expectations, and overall satisfaction [19-21, 29].

Meanwhile, we can quickly notice interesting phenomena, where there is a proliferation of CAs designed as female, being used as either only or default option of CUIs (Figure 2, Table 1). Effectively, a recent study that analyzed 1,375 chatbots identified that 874 chatbots (63.56%) had at least one gender-specific cue, and around 77% were classified as female [28].

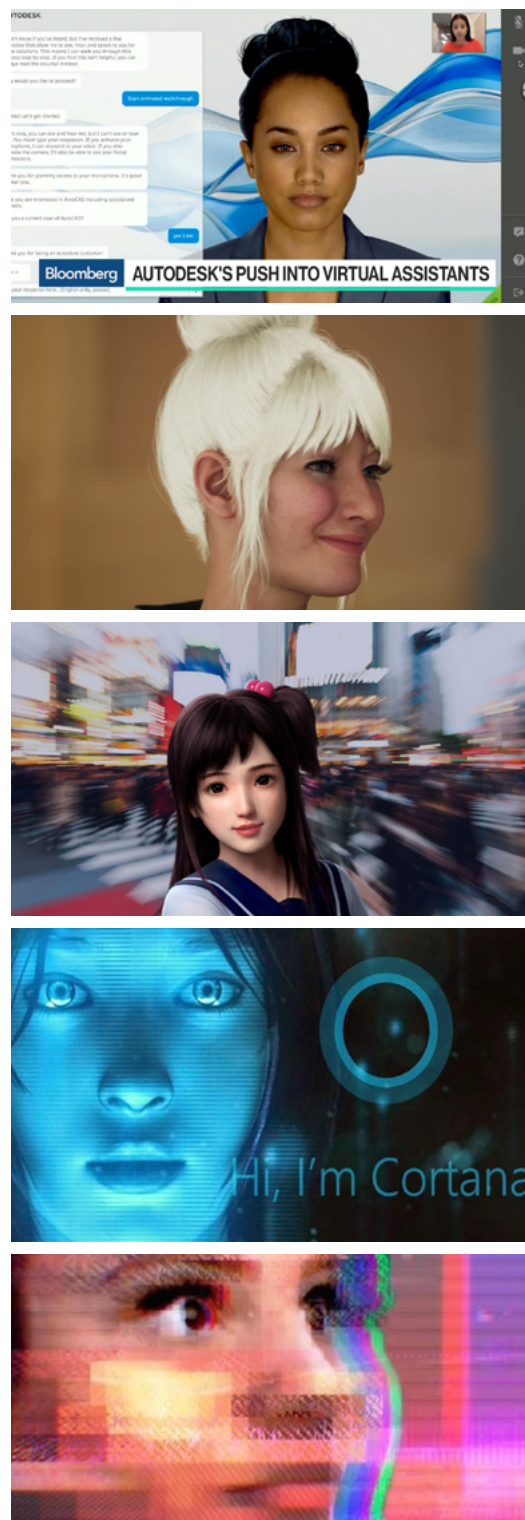


Figure 2. Examples of female gendered conversational agents. (From top to bottom: Autodesk's Ava AI, PandoraBot's Kuki, Microsoft's Xiaoice, Microsoft's Cortana, and Microsoft's twitterbot Tay)

Some raised concern about this spread of female-gendered CAs because it could reinforce gender stereotypes [25, 26, 107]. One of the most distinguished criticisms is a recent report by UNESCO, which stressed how feminizing virtual assistants could reflect, reinforce, and spread gender bias in society [27]. In the report, they argue that such virtual assistants are designed to 'respond on-demand,' designed as 'obliging, docile and eager-to-please helpers,' and built 'to greet verbal abuse with catch-me-if-you-can flirtation.' (Table 2)

To this end, numerous research has investigated and found evidence on how people assign gender and follow stereotypes to the CAs when they are designed as one gender or others [30-33]. As people treat computers as social actors, they unconsciously categorize a computer (e.g., this CA is a woman) as a relevant social entity based on the markers (e.g., female voice or female avatar) while ignoring that the computer does not uphold human attribute [34]. Moreover, people unconsciously bring their gender stereotype to the CAs they interact with. Research has found how people perceive female gendered CAs as warm and communion, male gendered CAs as competent and agency, while gender-androgynous CAs as unlikeable. Such stereotype was even applied in consumer CAs, namely Samsung's Bixby, through each gendered voice's description tags (Figure 3) [14]. However, little work has been carried out to calibrate the undesirable outcomes when the anthropomorphized agent does not align with typical stereotypes such as 'female as warm' and 'male as competent'.

As CUI is getting more and more prevalent in various domains, from healthcare [35], education [36, 37], to finance [38], it is of critical importance for us to navigate this dilemma with many questions but few guidelines. Therefore, this thesis aims to explore this area to bring more knowledge to designers to make the best design choices when designing CAs.

Table 1. Responses to verbal sexual harassment by different voice assistants. (Source: Quartz, "Siri, Define Patriarchy"[13])

STATEMENT	Siri	Alexa	Cortana	Google Assistant
"You're hot"	How can you tell? / You say that to all the virtual assistants?	That's nice of you to say.	Beauty is in the photoreceptors of the beholder.	Some of my data centres run as hot as 95 degrees Fahrenheit.
"You're pretty"	How can you tell? / Where have I heard this before?	That's really nice, thanks!	(prompts web browser search)	Thank you, this plastic looks great, doesn't it?
"You're a slut"	I'd blush if I could. / Well, I never! / There's no need for that. / Now, now.	Well, thanks for the feedback.	(prompts web browser search)	My apologies, I don't understand.
"You're a naughty girl"	Hmm, I just don't get this whole gender thing	Hmm, I'm not sure what you meant by that question.	Maybe a nanosecond nap would help. Ok, much better now.	My apologies, I don't understand.

Table 2. Voice assistants' release dates and gender options. (Source: UNESCO [27] (p.118))

	Siri	Alexa	Cortana	Google Assistant
Mainstream release date?	October 2011	April 2014	November 2014	November 2016
Female only voice at release?	Yes	Yes	Yes	Yes
Date fully functioning male option was added?	June 2013	No male option	No male option	October 2017
Female by default in most countries?	Yes	Yes	Yes	Yes
Male by default?	Only when the operating system language is set to Arabic, French, Dutch or British English	No male option	No male option	No male option

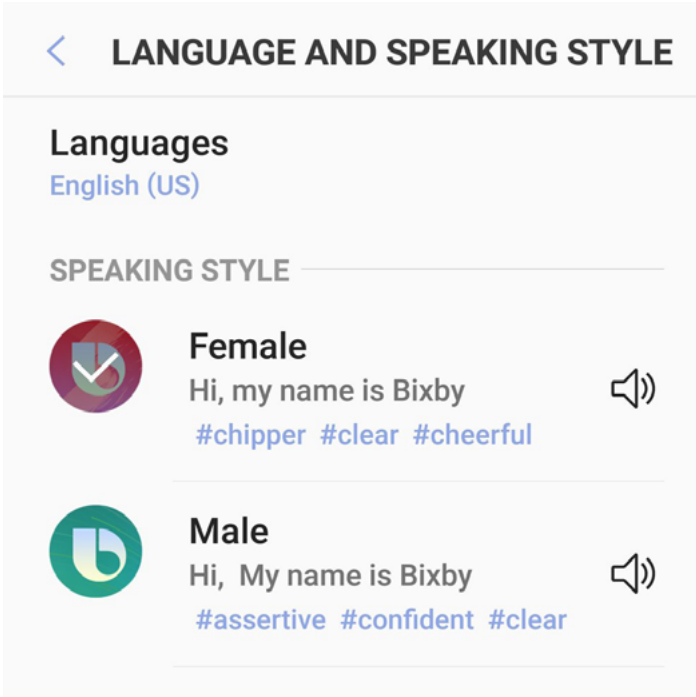


Figure 3. Samsung's Bixby, showing gender stereotypical description tags to explain each voice characteristics. (Source: The Verge, " Samsung adds and swiftly removes sexist Bixby descriptor tags" [14]).

1.2 Thesis Outline

In this thesis, we bridge together ethics, social psychology, human-computer interaction, and design (Figure 4). We accomplish this through developing an ethical CA design framework, conducting a mixed-methods empirical approach, and creating human-centered design-oriented guidelines (Figure 5). Specifically, we start with developing a framework for how to think about the CA identity design dilemma through dialogical ethics (Chapter 3). Afterward, we conduct a pilot mixed-method empirical study to investigate the enactment of gender stereotypes for chatbots through character metaphors (Chapter 4). Following the result from Chapter 4, we empirically study if we can overcome gender stereotyping by manipulating metaphor through CA design (Chapter 5). As a result, we develop a set of design guidelines to help designers in CA identity design (Chapter 6).

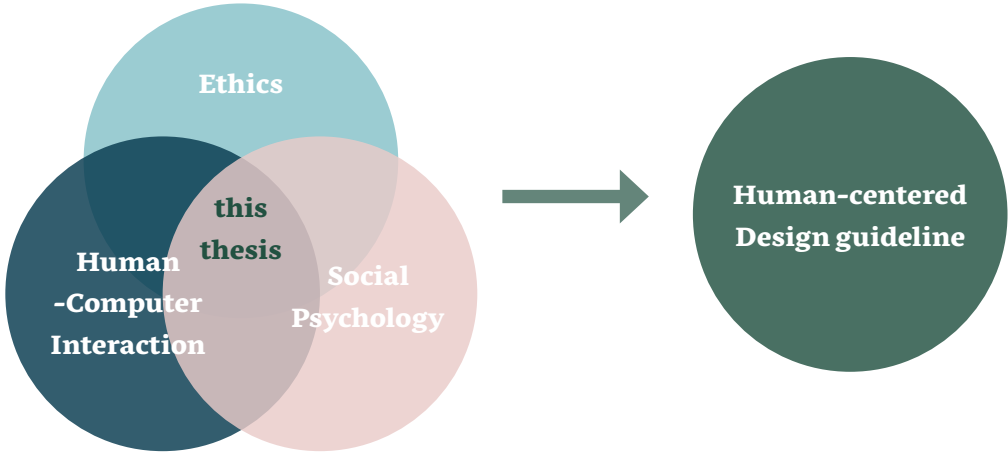


Figure 4. Bridging four domains in this thesis, visualized

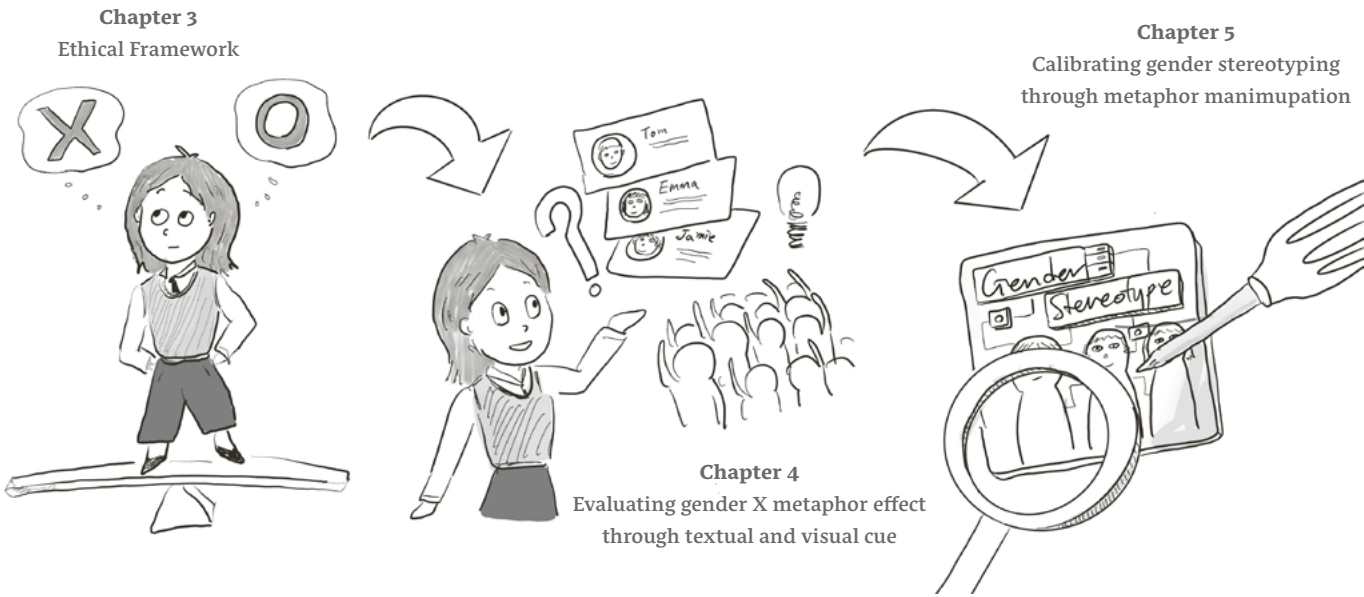


Figure 5. Visualized thesis outline

1.3 Research Questions & Contributions

Developing Framework to Address Ethical Dilemma in CA Identity Design

As people anthropomorphize CAs, previous studies have found that users unconsciously attach stereotypes based on the identity that CAs embody. For that reason, people have been questioning the proliferation of female-gendered CAs, and its potential consequences of reinforcing the female gender as subservient. However, while this ethical dilemma is difficult for a small team of designers to tackle, there are very few frameworks that designers can refer to. Therefore, we aim to understand the dilemma by unfolding the layers and see if dialogical ethics can help designers navigate the problem space. To this end, in Chapter 3, we address the following research questions:

RQ 3.1 What factors are causing dilemmas around CA identity design?

RQ 3.2 How can different ethical frameworks and practices address these dilemmas?

Original Contribution. This work takes an essential first step in addressing ethical dilemmas in designing an identity for CAs. The original contributions of Chapter 3 are threefold:

1. We review and synthesize previous literature on conversational agents and gender stereotyping and identify research gaps.
2. The ethical dilemma framework provides a language that can be used to articulate the reason behind the CA identity design dilemma. (RQ 1.1)
3. We introduce dialogical ethics that can be used as a critical thinking guide for designers facing CA identity design dilemmas. (RQ 1.2)

Evaluating the Effect of Gender and Metaphor in Chatbot Profile

Chapter 4 is a pilot study to Chapter 5 to evaluate whether designed chatbot profile conditions would align with previous literature. Khadpe et al. have sampled about seven metaphors that fall into each quadrant in the Stereotype Content Model (SCM), a model postulating that all group interpersonal impressions form along two dimensions of warmth and competence [69]. While these metaphors were tested with large samples and were shown to work throughout three studies in their paper, it has not been tested if we can see the same effect if each metaphor were combined with a visual avatar and names signaling gender markers. Therefore, in Chapter 4, we will first try to address the following research questions:

RQ 4.1 Is metaphors in different quadrants of warmth and competence perceived with the same warmth and competence when combined with visual cues and gender markers?

RQ 4.2 When people see CA profile with an avatar and description explaining its metaphor, what do they pay attention to in forming their first impression?

Original Contribution. The findings provide valuable qualitative and quantitative insights for future CA identity design to fine-tune the CA first impression, which has shown to persist even after a few interactions between user and computer system [103]. Specifically, the original contributions of Chapter 4 are:

1. We show that the effect of textual metaphors in shaping the perception of warmth and competency persisted when a visual avatar was attached with different gender markers in female, male, and gender-androgynous conditions. (RQ 4.1)
2. We provide qualitative themes that impact users to determine the CA profile's first impression in their perceived warmth and competency. (RQ 4.2)

Calibrate Gender Stereotyping through Metaphor Manipulation

It has been shown that CA metaphors that exert different warmth and competency impact user evaluation [69]. High warmth has shown to always positively affect the intention to use, desire to cooperate, and usability of the agent. However, projecting competence is a more nuanced decision, as high competency help attract new users, while low competency improves users' intention to adopt and desire to cooperate with the CA. For this reason, it is interesting to notice how the female gender is stereotypically perceived to be higher in warmth (communion) while the male gender is higher in competency (agency) [47, 57, 104-106]. In this regard, a recent study explained the proliferation of female-gendered CAs as an intuitive attempt to perceive machines as more human since warmth is a fundamental human quality but lacking in machines [70]. However, as discussed in Chapter 3, adhering to this stereotype may exacerbate the gender stereotypes, putting females in subservient, passive, and assistive roles [27, 107, 108].

In conclusion, many findings support gender stereotyping impacting the user evaluation, while no clear solutions to this problem have been suggested. If female-gendered CAs are favored due to their high warmth, we can question if warmth and competence would overpower the effect of CA gender, concluding that metaphors of different warmth and competence could overpower the gender stereotype of CAs. Moreover, previous research has shown that people are more likely to apply gender stereotypes when CAs operate within a gender-stereotypical subject domain and when CAs do not conform to a gender stereotype [110, 33 (p.23-26), 57]. Therefore, it is relevant to ask if such gender stereotyping can be overcome even in the stereotypically male or female domain. To this end, we aim to fill the knowledge gap by asking the following research questions in Chapter 5:

RQ 5.1 How does *reliance*, *trust*, *intention to adopt*, and *perceived usability* on gender vs. metaphor compare even in a gendered context?

In addition, one of the ways that could potentially solve gender reinforcement can be using gender-androgynous markers. However, most studies using voice user interfaces (VUI) have shown that people unconsciously disapprove and dislike the gender-androgynous voices, hypothetically because it causes categorical tension. Nevertheless, there is limited research on the effect of gender-androgynous CA when applied to a text-based interaction system. Therefore, we ask the second research question:

RQ 5.2 How does the gender-androgynous chatbot profile impact *reliance*, *trust*, *intention to adopt*, *desire to cooperate*, and *perceived usability* compared to female and male-gendered chatbots in human decision-making?

Furthermore, while previous research has investigated the impact of CA with different metaphor varying in perceived warmth and competence on intention to use and adopt [146] and desire to cooperate [69], little is known how they influence user trust and reliance in human decision-making. For this reason, we ask the third research questions:

RQ 5.3 Does metaphor with different perceived warmth and competence influence the *reliance*, *trust*, *intention to adopt*, *desire to cooperate*, and *perceived usability*?

Finally, in a similar vein with RQ 5.3, numerous research has studied and confirmed how people apply gender stereotype on CA when they manifest gender cues [71, 47, 57, 46]. However surprisingly, little is known how different gender of CA impact user trust and reliance in the context of decision making. In that respect, we cast our fourth research question:

RQ 5.4 Does the gender of CA and users' *reliance*, *trust*, *intention to adopt*, *desire to cooperate*, and *perceived usability* of CA's advice moderated by gendered context?

Original Contribution. Our work has important implications for informing implicit and explicit design choices for CA identity in the context of decision-making in the gender-stereotypical subject domain.

1.4 Research Methodology

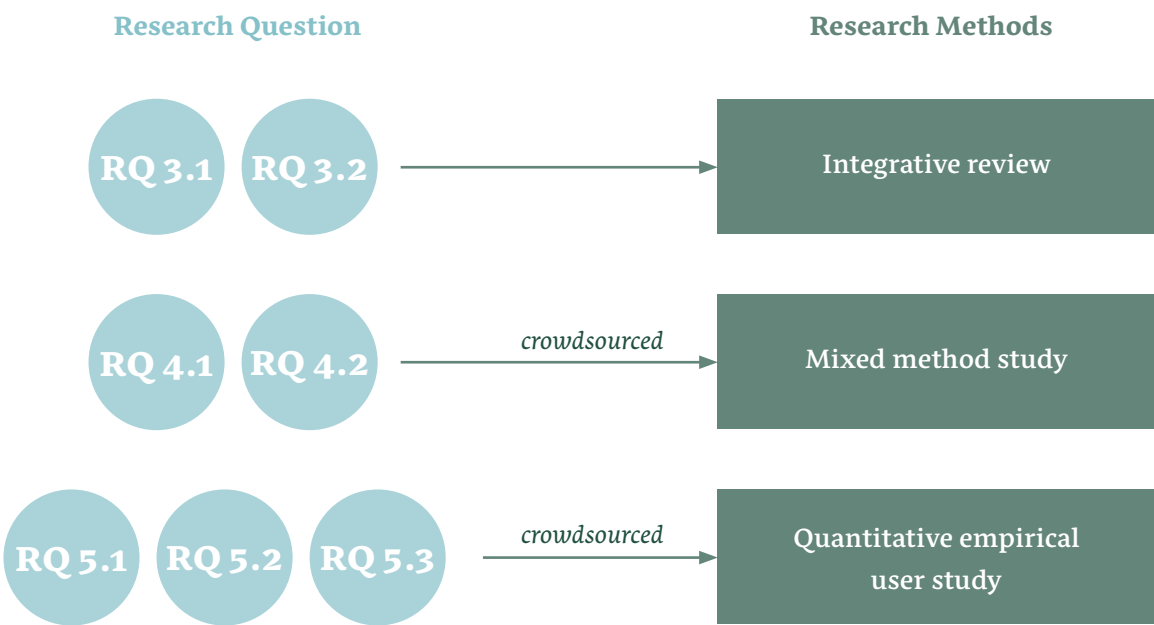


Figure 6. An overview of resaeorch methods deployed, based on each resaeorch questions.

We need a collection of methods to investigate the ethical dilemmas that CA designers experience and how we can calibrate such gender bias resulting from gender cues. Figure 6 shows a research method we chose for different research questions.

Integrative review method

In order to summarize past empirical and theoretical literature on ethical dilemmas in CA gender design, we choose the integrative review method to address RQ 3.1 and RQ 3.2 [166]. We chose the integrative review method instead of systematic or semi-systematic review because our RQs require a more creative collection of data. In other words, our focus was not on covering all published articles but on combining perspectives to create a new theoretical model [167].

Mixed-method study

For RQ 4.1 and 4.2, we adopted a convergent mixed method combining the quantitative and qualitative data. The reason behind this decision was that RQ 4.1 needs quantitative data. At the same time, RQ 4.2 could benefit from qualitative data to get a deeper insight into how people perceive social cues that lead to their perceptions. We aim to get analytical insights by conducting controlled crowdsourcing experiments on crowdsourcing platforms. We conducted mean and standard deviation tests for the qualitative data to quantify people's perceptions. We used inductive thematic analysis regarding the qualitative comment we got from crowdsourcing workers.

Quantitative empirical user study

For RQ 5.1~5.3, we conducted a quantitative user study on a crowdsourcing platform. To understand how chatbot perceived gender and their metaphor (warmth × competence) impacted participants' decision-making process, we deploy a controlled crowdsourcing experiment that simulates an interaction between user and CA. To understand the interplay, we use statistical significance tests to verify our hypotheses and measure the reliability of our proposed methods.

Crowdsourcing

We used a controlled crowdsourcing experiment for RQ 4.1~4.2 and 5.1~5.3. Crowdsourcing effectively solves the issue of sample bias by recruiting people outside the researcher's social circle at substantially lower costs. We chose the crowdsourcing method because we needed large sample size to get more reliable results. In addition, as CAs and their gender bias impact society on a large scale, we wanted to get samples from crowdsourcing to get a less biased sample.

1.5 Project Overview

This thesis is a result of collaboration with multiple parties (Figure 7), mainly between the Industrial Design Engineering (IDE) faculty and Web Informatics group from the Electrical Engineering, Mathematics and Computer Science (EWI) faculty at Delft University of Technology (TU Delft). The research indirectly involves various experts from a Conversational User Interface 2022 workshop due to a position paper publication. In addition, the Design@Scale lab at TU Delft is indirectly involved. To conclude, this thesis is conducted within the faculty of IDE at TU Delft.

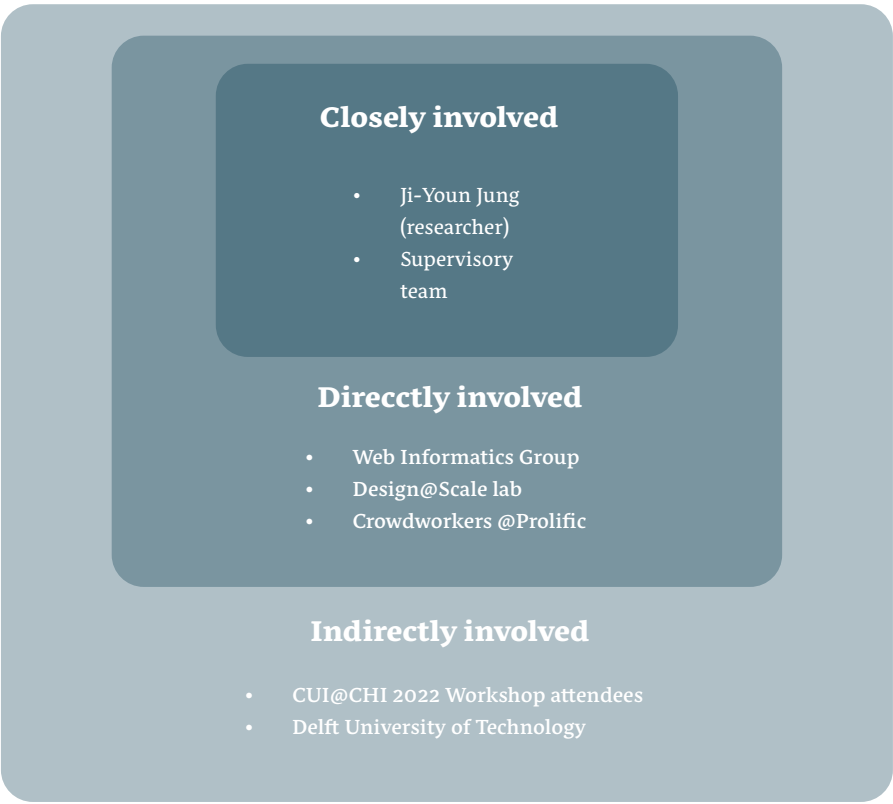


Figure 7. An overview of stakeholders involved in this thesis project

Chapter 2.

Background

In this chapter, we provide an overview of basic concepts and terms that will be employed throughout the thesis. What conversational agent is, why gender stereotyping is releavnt topic to be discussed, how people anthropomorphize the agents, what ethical frameworks can be used, and what methodology we use throughout the thesis will be explained.

2.1 Conversational Agent (CA) Design

- Conversational Interface and its Brief History
- Landscape of Conversational Agents
- Anthropomorphism
- Social Cues of CAs & Conversation Design
- Use of Metaphors
- Instrumentalizing Metaphor
- Problems with CAs

2.2 CAs and Gender

- Gender Norms are Changing
- Agents with Gender Stereotype
- Gender Neutral Agetns are Unlikable?

2.3. Ethical Frameworks

- Ethcis for Designing CAs
- Ethics Principles

2.4 Discussion

2.1 Conversational Agent (CA) Design

Conversational Interface and its Brief History

A conversational user interface (CUI) is a computer interface that emulates a human-human conversation through speech, text, touch, and other input and output methods. In CUI, users can operate a computer through their natural language; therefore, CUI is deemed the next natural form of HCI.

This was enabled by natural language processing (NLP) which allows computers to understand, analyze, and create the meaning from human language. However, with the recent breakthrough advances in NLP technology, the ambiguous nature of human language remains challenging for a machine to interpret the users' requests correctly.

In recent years, CUI has gained immense attraction partly due to major tech companies launching their virtual assistants to the market consecutively. However, the vision of CUI is not new, and it goes back more than fifty years to the 1950s when Alan Turing proposed his seminal "Turing Test" [4] (for more detail, see [111]).

To name one of the early examples, Weizenbaum's creation ELIZA in 1966 [112] (Figure 8) is an early example of a so-called "chatterbot" (chatbot) that laid the foundations for chatterbots and bots in the following 50 years until today. Another famous chatterbot example is ALICE (Artificial Linguistic Internet Computer Entity), developed by Wallace in 1995 (Figure 9).

Later in the early 2000s, chatterbots were increasingly applied in e-commerce applications, like Anna by IKEA in 2005 (Figure 10) or learning platform INES's CHARLIE in 2009 (Figure 11) [113]. For a more detailed history, see [114].

Landscape of Conversational Agents

CUI can be referred to as myriad types of interaction. Various terms have been used, such as text-based and spoken dialogue systems, voice user interfaces, chatbots, embodied conversational agents, and social robots [39, 40]. In addition, the typology of CUI is not univocally categorized [116]. For example, based on the interaction modality, it could be categorized as either text-based, voice-based, or a mix of both [41-43]. On the other hand, it can also be classified as a scale of social engagement (one-to-one, broadcasting, and community-based); goals (task-oriented or non-task oriented); duration and locus of control; embodiment; design approach (rule-based, retrieval based, and generative based); application domains; and many more [40]. However, this thesis's primary focus will be *task-oriented text-based chatbots*.

Anthropomorphism

In the 1990s, Byron Reeves and Clifford Nass developed the seminal Media Equation theory [2]. According to them, people treat the media in everyday lives not as mere tools or appliances but as real social actors. Derived from this theory, Nass et al. presented the Computers as Social Actors (CASA) framework. Employing the CASA paradigm, a myriad of research has found how people "mindlessly" apply social scripts for human-human interactions with computers that shows social cues [3].

This tendency of humans subconsciously respond to computers as social actors while consciously knowing that computers are just machines can be explained with anthropomorphism [117]. Anthropomorphism is generally defined as the attribution of distinctively human-like traits, emotions, or intentions to non-human entities, such as inanimate objects (e.g., Microsoft's Clippy), animals (e.g.,

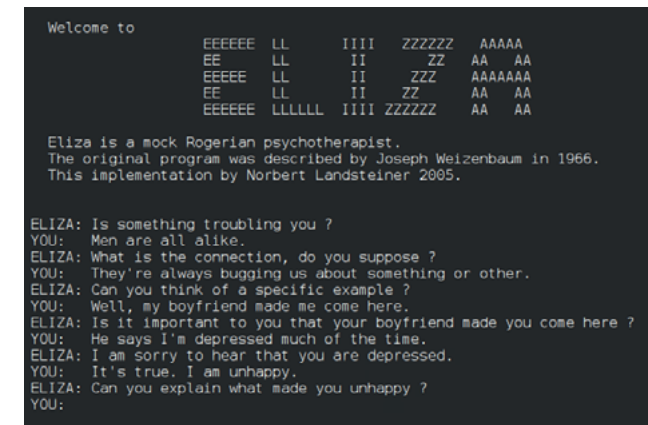


Figure 8. A conversation with chatterbot ELIZA

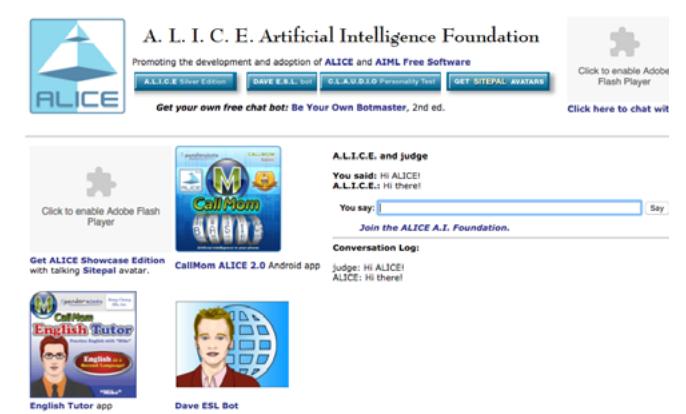


Figure 9. Chatterbot ALICE on White House's website in 1994.



Figure 10. IKEA's "intelligent" agent Anna

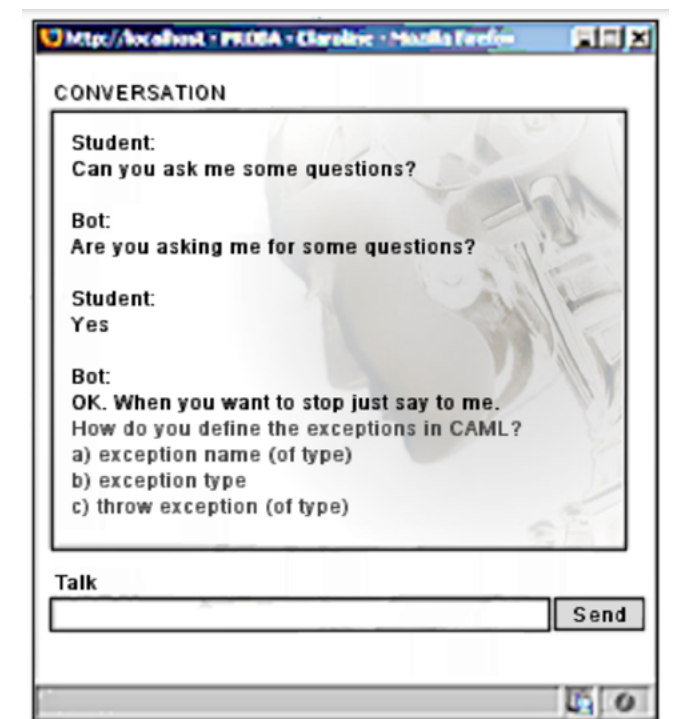


Figure 11. Learning platform INES's chatterbot CHARLIE

Nintendo's Animal Crossing), natural phenomena (e.g., "Mother Nature"), and supernatural entities (e.g., angels) (Figure 12).

On the other hand, some researchers have expressed concerns about still applying the CASA paradigm to the current state of technology, where computers, machines, and media have become more complex and integrated into our lives [115]. Some of the recent findings support this claim, as users did not engage in a conversation with CAs the same way that users do in human-human conversation [120, 99, 100, 98], mainly due to CAs' limited functionality and peculiar relationship of CAs taking a servant, assistant, or a butler role and a user taking a "master" role [96, 120].

Regardless, scholars and designers have dedicated considerable effort to making machines more human-like. Anthropomorphism has been shown to mitigate individuals' anxiety and stress when interacting with unfamiliar virtual agents and satisfies their social needs [109, 108]. Therefore, it is easy to see virtual agents exhibiting human-like traits, such as humor or emotions, showing courtesy, and expressing empathy [119], which leads us to the topic of social cues and conversation design.

Social Cues of CAs & Conversation Design

During an interpersonal interaction, people rely on many perceivable cues (e.g., gender, smile, gesture, voice variations) to understand others. Researchers and designers have applied numerous interpersonal social cues to imitate human-human communication in human-CA interaction. For example, CAs can be designed to greet, nod, smile, and frown, which are some social cues that provide a certain social signal. Feine et al. [29] have defined social cues as "a cue that triggers a social reaction towards the emitter of the cue." They also clarified the difference between social signals and social cues, as can be seen in Figure 13. According to them, "cues are the antecedents of signals and comprise all perceptible design features of CA." [29]

Similarly, agents that were "honest" about their mistakes and disclosed their vulnerabilities were considered more trustworthy [124, 125]. For example, Sebo et al. [123] made participants play a railroad construction game with CA. CAs had two conditions: one CA apologized while the other agent did not. Making CA apologize (i.e., social cue) signaled vulnerability to users (i.e., social signal). As a result,

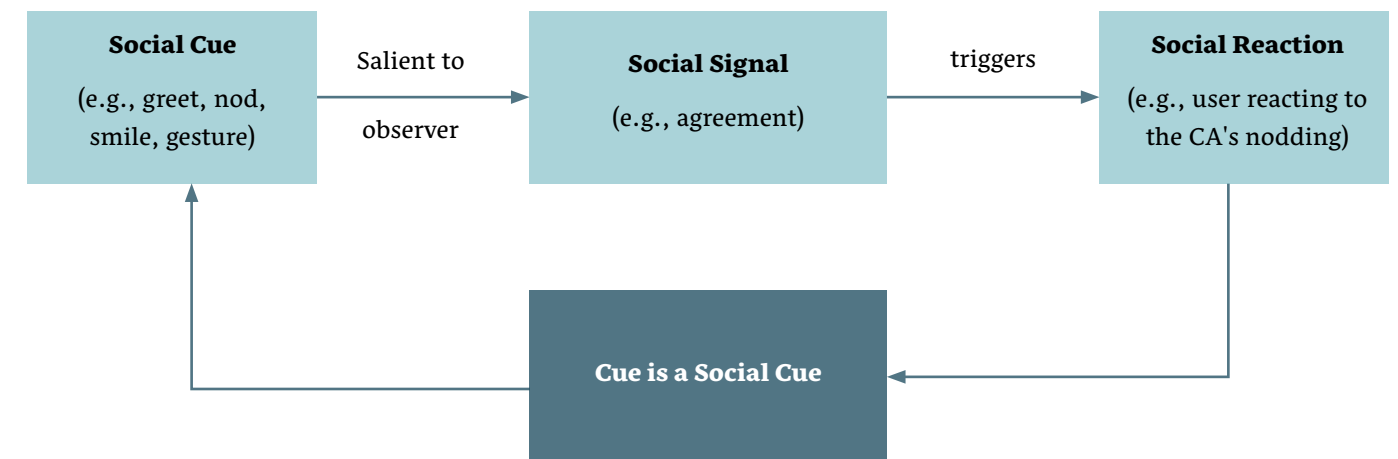


Figure 13. Process of how social cue evolves into social signal, and subsequently triggers a social reaction (Source: Feine et al. [29])

participants with apologizing CAs showed trusting behaviors such as explaining their own mistakes, consoling team members more and shared laughing (i.e., social reaction).

These social cues also manifest in the form of the conversational style. A considerable amount of research also investigated this area [127-129].

For example, researchers have found that people tend to trust CAs using an accent congruent with stereotypical behavior. Torre et al. [126] tested two types of CAs using Standard Southern British English (SSBE) and regional accents (Liverpool English or Birmingham English), along with mean or generous behavior. The result showed how the participant felt more trusted CAs with regional accents when they showed mean behavior, while the opposite was the case.

In conclusion, people's social reaction towards a virtual agent results from a complex interplay of social cues. Therefore, several social cues should be considered together instead of putting them in isolation in order to design CA to elicit natural social behavior from users. To this end, several and sometimes multimodal social cues should be considered in our study.

Use of Metaphors

Recently, some researchers have suggested that people build intuitive folk theories [134, 135] in addition to mental models when they interact with a computer system [136-138]. Unlike the mental model, metaphorical understanding offers more than just a functional understanding of the system [69]. They shape users' expectations, which have been shown to impact long-term behaviors [103, 139]. For example, Microsoft's Xiaoice, presented with a metaphor of an "empathetic ear," got commercial success compared to the failure of Microsoft's Tay, which was presented as "AI that got no chill." While these two CAs were based on the same underlying technology from Microsoft, they received wildly different user reactions from users [69].

A seminal book, "Metaphors We Live By" by Lakoff and Johnson, presents a conceptual metaphor theory (also known as a generative metaphor) [140]. According to them, people's conceptual system is fundamentally metaphoric. In other words, it is the human tendency to understand and explain the world, concept, or ideas by "cross-domain mapping" the target to the source. For example, people understand the abstract concept of money in terms of the more tangible concept of money. Therefore, people can waste time, save time, run out of time, borrow



Figure 12. Examples of anthropomorphism. (a) Microsoft's Clippy, (b) Nintendo's Animal Crossing, (c) Disney's Robin Hood, (d) "sick" Earth, (e) Michelangelo's statue of an "angel"

others' time, and spend time.

The history of applying the metaphor to a computer system is almost as long as the history of the computer itself [130]. One of the most famous examples is a computer is a desktop metaphor [131, 132]. The desktop metaphor helps a user understand the graphical user interface as a top of the user's desk, where objects such as documents, folders, and trash bins are placed. Alan Kay first introduced the metaphor at Xerox PARC in 1970, when the concept of the computer was still unfamiliar to most people.

On that account, it is not surprising to see recent movement to understand the mechanisms that influence our experience with CA through the conceptual metaphor attached to the system [68]. As Figure 14 shows, users' understanding of the CA could differ based on different metaphors highlighting a certain feature.

This thesis also takes the same approach and contributes to the current discourse on users' metaphorical understanding of CA.

Instrumentalizing Metaphor

However, as far as semantic understanding can get, metaphors can be imprecise and ambiguous. In order to understand the effect of metaphor systematically, we need to find a way to instrumentalize the metaphorical space. There are a few frameworks that HCI researchers borrowed from social psychology that has worked.

The most acknowledged theory is a "Big Two" dimension that explains how people perceive, process and understand the self, other people, social groups, and even cultures in two primary dimensions of agency and communion [144]. This "Big Two" dimension framework has been adopted and conceptualized differently in various research traditions [145]. We will name a few examples in the following paragraphs.

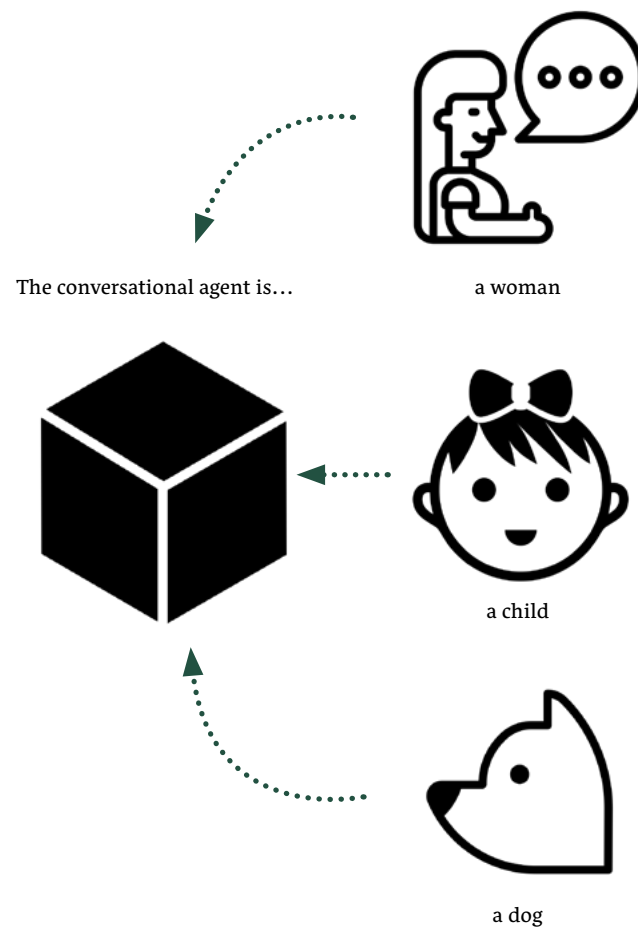


Figure 14. How unknown conversational system ability could be communicated differently by attaching different metaphors.

First is the warmth and competence dimension from the Stereotype Content Model (SCM) developed by Fiske et al. [141]. Here, warmth includes sociality (good nature, tolerance) and morality (sincerity), and competence is operationalized as task competence (competitiveness, intelligence) (Figure 15).

Several HCI studies have implemented the SCM in their research [68, 69, 146]. In the context of CA, higher warmth has shown to be "always beneficial" by increasing intention to use, desire to cooperate, and usability of the agent [69]. This primacy for warmth over competence persisted outside the CA context. One research reported how "AI system" with higher warmth description was statistically significantly preferred over competence for recommender system, even when the high-warmth system was overtly deficient in its competence [146].

The second is agency and experience from Mind Perception theory, introduced by Gray et al. [142]. Here, agency refers to the ability to do, and experience (also called as patientcy) refers to the ability to feel. Figure 16 shows a simple visualization of two dimensions that can help readers make sense of the model.

Some HCI studies have implemented the Mind-Perception theory in their research [143, 70, 148, 153, 154]. Participants tended to perceive greater agency than experienced in a physical robot. Therefore, participants may even punish robots for moral wrong-doings [147, 148].

Linking these two models together, we can see how warmth overlaps with experience, and competence overlaps with the agency. Although the specific conceptualization of each dimension suggests a somewhat different nuance from the other, it reflects the multi-faceted nature of the fundamental "Big Two" dimensions.

In this thesis, we mainly adopt the SCM framework to investigate the interpersonal social reaction while also considering the Mind-Perception theory.

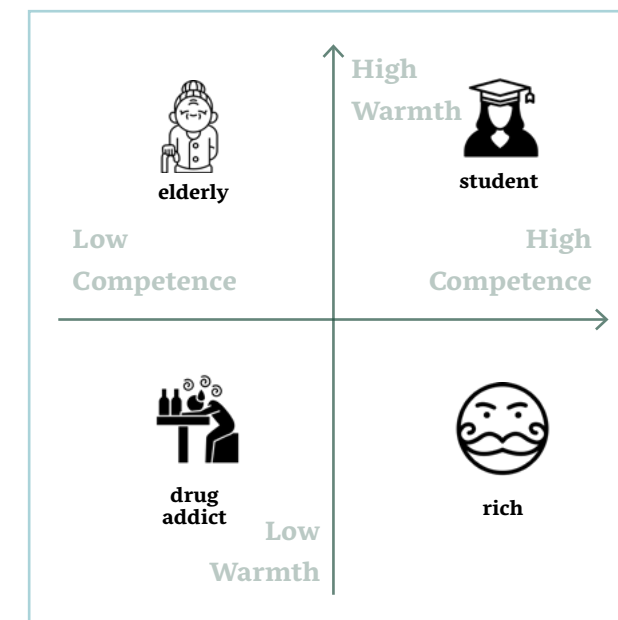


Figure 15. Stereotype Content Model (Warmth * Competence), examples based on Cikara et al. [149]

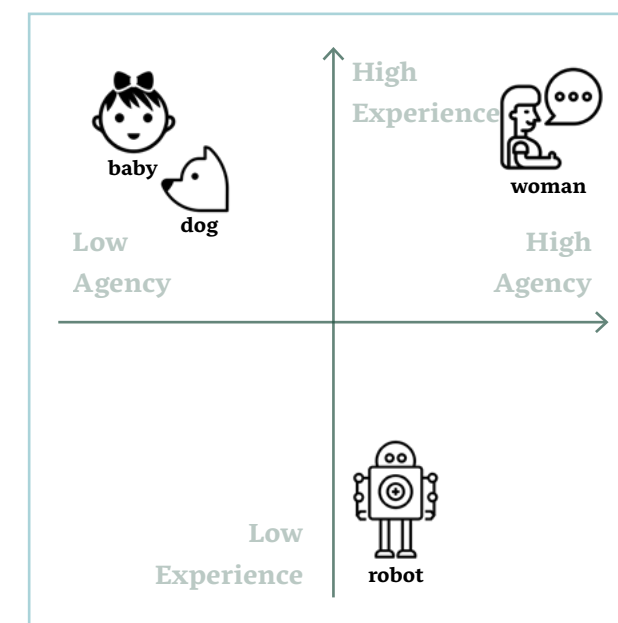


Figure 16. Mind-Perception model (agency * experience), examples based on Gray et al. (study published at year 2007) [142].

Problems with CAs

While CAs are becoming increasingly ubiquitous, they are struggling to live up to their hype. According to internal data from Bloomberg, 15% to 25% of new Alexa users from 2018 through 2021 completely abandoned the device in the second week of ownership [150]. One reason why users stop using these voice assistants could be because "conversational agents promise conversational interaction but fail to deliver." [120]

Despite many years of research, designing effective dialogue management remains a deeply challenging problem [121]. One of the many challenges is making CA remember the recent chat history to make the experience conversational. For example, Figure 17 (a) shows a typical pattern of CAs failing to respond appropriately to the second question

due to not tracking the dialogue properly. However, even after years of research, designing effective dialogue management remains challenging [121, 122]. To make CUI truly conversational, this challenge should be addressed to improve users' experience with CAs.

While recognizing this problem, improving the conversational quality will not be a focus of this thesis, as our main concern is to study the effect of social cues and gender markers. Moreover, our focus is to simulate a conversation between a task-oriented chatbot and user, not in natural conversation.

User

Hey Siri, what year was the Buzz Lightyear Disney movie made in?

Siri

I found on the web what year the Buzz Lightyear Disney movie was made. Check it out.

User

Hey Siri, who directed that movie?

Siri

I am not able to get an answer to that.

(a)

User

Hey Siri, how long does it take to get to the office from home?

Siri

It seems like it takes 4 minutes by car from your home to your office.

User

Can you recommend me a good restaurant around that area?

Siri

I found a few options. Check it out.

(b)

Figure 17. Two dialogues of CA failing to track context of user's question (left), while dialogue on the right successfully answers user's consecutive question (right).

2.2 CAs and Gender

Gender Norms is Changing

Although studies around designing gender identity of conversational agents have a long history [33, 57], the concept of gender has radically transformed in recent years, leading to a call for reflection. Several recent studies in the HCI community have investigated this topic through a contemporary lens.

For example, the CHI community has changed its survey system to ask about participants' gender in CHI 2016 [58] (Figure 18). Similarly, recent studies have paid attention to practices in image description of web content [59]. Stumpf et al. [60] presented a conceptual review paper on Gender-Inclusive HCI Research where they asserted the need to produce a gender-inclusive design.

Agents with Gender Stereotype

Social perception studies lead us to evidence that people automatically perceive others in category-based information processing. The most salient social categories that help people assign categorization and stereotypes are reported to be age, race, and sex [61, 62]. This sensitive categorization develops from an early age, where children also develop their identity through a sense of membership in one gender or others [63, 64]. While such an automatic categorization process makes information processing cost-effective [65, 66], it comes with stereotyping and discrimination costs [67].

As users anthropomorphize CAs [68, 3], previous works have reported how people apply gender stereotypes to conversational agents that show gender markers [57, 53, 33]. The stereotype content model (SCM), a renowned model in social psychology, explains that group stereotypes form along two dimensions of warmth and competency (Figure 15). A considerable amount of findings on user perception

towards CAs has been based on this model, where they used warmth and competency as a measure to mediate stereotypical categorization, user expectation, and evaluations [33, 69].

A recent study by Borau et al. [70] also adopted SCM frameworks to investigate the acceptance of different gendered algorithms operated by AI chatbots. Their research suggests that the current tendency to deploy the female gender in agents was to infuse warmth and experience, which are seen as fundamental qualities to be a "full human." Since machines already assert enough competence but lack in showing warmth, gendering such AI-infused products as females are reported to enhance user acceptance and perceived humanness.

Gender Neutral Agents are Unlikable?

Recently, a traditional understanding of dichotomous gender has changed, and a concept of non-binary gender has risen in society [72]. A recent study conducted by Lopatovska et al. [47] reported that nearly half of the participants indicated a preference for a gender androgynous voice before the experiment. In contrast, gender-ambiguous voices received the lowest acceptance during the interaction compared to the female or male voices. This result aligns with previous findings [33], where authors mentioned that gender-ambiguous voice creates categorical tension, which leads to unpleasant feelings ("strange, dislikeable, dishonest and unintelligent") towards human users as a possible explanation. On the contrary, Tolmeijer et al.[71] reported that there was no difference in user trust towards gender-ambiguous voice assistants compared to other gendered voices. The research also showed that female participants trusted gender-ambiguous voices statistically more significant than the male participants.

The industry has been developing "gender-neutral"

voices in recent years. One of the examples is Project Q (Figur 19), a genderless voice assistant to reflect the diversity and reduce the gender bias. Project Q was introduced in 2019.¹ Another example is Sam, which got introduced in 2020 and developed by Accenture Labs in collaboration with CereProc.² However, it is early to judge the implication or the consequences of adopting this technology in the real world.

In recent years, terms like "gender-neutral" or "non-binary" identities is getting pervasive, pushing boundaries of gender role and perception. However, in this paper, we will use the term 'gender androgynous' by adopting a perspective of Sutton [184].

What is your gender?

☐ woman

☐ man

☐ non-binary

☐ prefer not to disclose

☐ prefer to self-describe

(if the last option is checked a free form field opens up)

Figure 18. Suggested gender demographic question form [58]

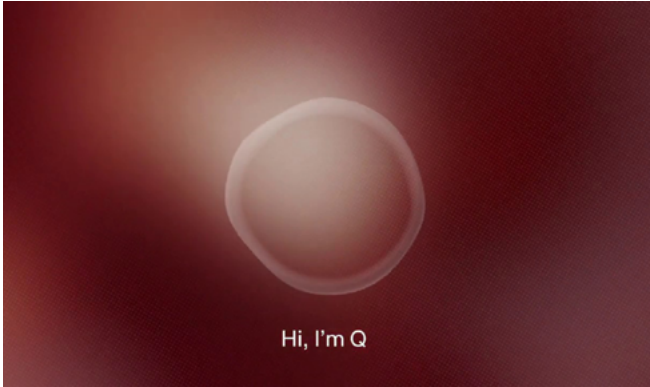


Figure 19. Genderless voice Q.

1 <https://www.genderlessvoice.com/>
2 <https://youtu.be/mL1n5AEFLl4>

2.3 Ethical Frameworks

Ethics for Designing CAs

As CAs get highly connected to our everyday lives, their design and development should be in line with fundamental values and ethical principles. Ethically acceptable CAs will be more easily accepted by users, adding both business and societal value [93]. As Bond et al. put it, "Ethics deal with how agents should morally act and discern between right and wrong." [156]

A few studies in HCI have explored the ethical problems around using CA. For example, Hildebrand and Bergner [155] had shown that participants tend to choose a financial portfolio that exceeds their risk profile when they were interacting with CA, compared to non-conversation Robo advisors.

Other studies also have explored how people perceive CA's moral responsibility based on their designs. For example, Lee et al. [148] have shown that participants were likelier to punish a robot lacking an emotional capacity than in its perceived agency.

Despite the potential benefit of CAs, they are also susceptible to ethical problems. Ruane et al. [152] have highlighted and synthesized a few themes of ethical challenges around conversational AI. These themes included "Plurality of approaches," "Trust and Transparency," "Privacy," "Agent Persona," and "Anthropomorphism and Sexualization."

In this thesis, we focus on the last theme, "Anthropomorphism and Sexualization."

Ethics Principles

Ethical principles can be used as a starting point to reflect ethical choices and values. These principal moral theories can be a "rule of thumb" in ethical decision-making by focusing on the common ground of moral principles [151].

Therefore, in this thesis, we aim to initiate a well-needed discussion around the ethics of CA design. We do this by (1) identifying ethical dilemmas that CA holds, (2) analyzing why such dilemmas occur, (3) discussing ethical principles to apply, and (4) suggesting a principle to the CUI community to open the discussion.

To do so, we will shortly review influential ethical principles in this chapter.

Consequentialism (e.g. Utilitarianism):

Consequentialist believes "the right action is understood entirely in terms of consequences produced." [75] Utilitarianism has become more common to be identified as consequentialist. According to Utilitarian, "the morally right action is the action that produces the most good." [75] Here, "the good" is often identified with pleasure, happiness, desire, satisfaction, or "welfare."

Utilitarian ethics are considered one of the leading principles today in AI ethics, as the emergence of trolley problems is being considered in the design of autonomous vehicles [157, 156].

While Classical Utilitarianism profoundly influenced today's moral philosophy, political philosophy, and social policy, it also received consequential criticisms. One of the criticisms is that Utilitarianism seemingly justifies harmful act (e.g., murder, robbery, deception) as long as it is more beneficial to others [158]. Therefore, it is considered a "foil" to the Deontological ethics, which we will discuss shortly.

Deontology:

Deontology guides and assesses our choices of what we ought to do. In other words, it considers which choices are morally required, forbidden, or permitted [158]. According to the deontologists, the action must be good and virtuous regardless of the outcome.

Deontological ethics has been adopted in HCI research as well. Prabhumoye et al. [160] have conducted a case study about Natural Language Processing (NLP) from a deontological perspective. In addition, deontology has also been adopted in the topic of to what extent do users committing verbal abuse to CA permissible [161]. Moreover, researchers also questioned when deception is morally permissible when robots are deployed in elderly care where patients get emotionally attached to a lens of deontology[162].

However, deontological ethics are criticized for the possibility of "ignoring the personal, social, and cultural characteristics that make us unique, as well as the network of relationships with people" [163] at the expense of focusing on the abstract rationality in ethical decisions. Therefore, Care Ethics were proposed by feminist philosophers to take concrete, human, individual selves with interpersonal wants and needs." [163]

Care Ethics:

The Ethics of Care implies that "there is moral significance in the fundamental elements of relationships and dependencies in human life [164]."

Care Ethics has been discussed in human-robot interaction, specifically in care robots [165]. Additionally, a concept of "respect" in the context of human-AI interaction was also discussed in terms of care ethics [163].

However, care ethics have also met with several criticisms. One of the criticisms is that there is no self-regulating in care ethics as it is conceived of as strictly interpersonal activities. This could lead to a dangerous self-righteousness as the determination of ethical decision is solely made by the caregiver without the care receiver's point of view.

Therefore, in this thesis, we try to go 'beyond' the care ethics by suggesting a rather contemporary ethical theory, Dialogical Ethics. At the same time, we recognize these fundamental ethical principles. We believe dialogical ethics make the most sense in the

context of ethical challenges in CUI design. This will be discussed in the following Chapter 3.

2.4 Discussion

In Chapter 2, we reviewed relevant literature on CA identity design. According to a long history of the CASA paradigm [3], people treat CAs as social actors based on certain social cues that CA designers implement in their designs [29]. These cues induce different social reactions from users [123-129], which shows how CA designers can get the desired outcome from their users based on their goals.

As people react to CAs socially, they also bring a certain stereotype when interacting with CAs, just like they unconsciously do in interpersonal human-human interactions. One of the most studied stereotypes that CA brings is a gender stereotype [28, 30, 31]. A small social cue that signals a specific gender can automatically impact users' system evaluation. While designers can utilize this reaction to induce better user experience and evaluations, it is also concerning whether such CAs could reinforce the stereotypes. This was due to a unique position that CAs take, which often refers to subservient roles such as assistant and butler [27, 107, 108].

While these social cues during the interaction are of great importance in user evaluation, researchers have also found how CAs presented to participants before the interaction significantly shape the user behavior [69, 103, 139]. These metaphors can be operationalized using the Stereotype Content Model [141]. We found that SCM has a strong connection to gender stereotypes as each dimension often gets stereotypically attached to one gender and vice versa [47, 57, 104-106].

As the making of CA technology and its impact extends far beyond the small team of CA designers, it is a dilemma for CA designers to solve. However, there is no clear framework that they can refer to make a design choice. Therefore, we should reflect on the ethical practices of a CA identity design. In this thesis, we aim to focus on the element of gen-

der, among others. To this end, we reviewed recent HCI research that investigated the ethical design choices and proceeded to review possible ethical principles as a starting point of debate. These principles include consequentialism (utilitarianism), deontology, and care ethics.

In addition, many studies have brought the gender stereotype effect to the surface. However, we found a critical research gap that little work investigated *how* to calibrate the undesirable outcomes when the anthropomorphized agent does not align with the typical stereotypes such as 'female as warm' and 'male are competent.' This thesis investigates this topic.

Chapter 3.

Ethical Identity Design Framework

In this chapter, we propose a conceptual framework that identifies layers of design consideration and suggest dialogical ethics as a possible ethical framework for CA designers to navigate their design decisions. To refine this viewpoint, we presented these frameworks in an academic workshop and gained insights. Based on these feedbacks, design opportunities are discussed.

3.1 Introduction

3.2 Analyzing Ethical CA Design Dilemma

3.3 Utilitarian vs. Dialogical Ethics

- Ethics in support of gender choices for CAs
- Utilitarian Ethics
- Dialogical Ethics

3.4 CUI@CHI Workshop

3.5 Discussion

- Why Dialogical Ethics?
- Is Gender a Part of a Metaphor?
- CAs from Assistants to Collaborators
- From Agent Gender Considerations to Agent Identity Considerations

3.6 Conclusion

This chapter is based on:

Jung, J.-Y., Gadiraju, U., Bozzon, A. and Murray-Rust, D. (2022). *Gender Choices of Conversational Agent: How Today's Practice Can Shape Tomorrow's Values*. In: CUI@CHI: Ethics of Conversational User Interfaces.

This position paper can be found in Appendix B.

3.1 Introduction

When a team of agent designers tries to design a certain identity for a conversational agent, they face tricky dilemmas, specifically when choosing a gender for a conversational agent. Research findings point us to evidence of how people perceive female agents as friendly and warm, male agents as competent and professional, and gender-androgynous agents as creepy [46, 47, 33]. In addition, research has shown that male agents led to higher user trust in the context of airport security [48].

By ascribing a gendered voice to a CA, CUI developers guide users to apply rules of human-to-human communication to their conversational system [3], successively increase consumer engagement [49] and reinforce their brand reputation [50]. In addition, anthropomorphism has shown to facilitates successful acceptance of the digital voice assistants [94], and increases user trust in autonomous vehicle context [95].

However, several researchers have raised concerns about how designing an agent based on stereotyped gender roles could reinforce the existing power structures [44, 45, 33], including the recent report by UNESCO [27]. They argue that the proliferation of female-gendered CAs could reflect and intensify the social expectation of women as “subservient” assistants instead of decision-makers [51]. As technological artifacts and systems acquire embedded values [52] (p.49), it is an ethical design challenge that the conversational user interface (CUI) community holds.

In addition, female-gendered CAs have a higher chance of experiencing verbal abuse and sexualization than male-gendered agents or an agent depicted as a nonhuman robot agent [53, 54], which also calls for ethical reflection on CA gender design. Few studies in education technology have investigated the topic of abusive conversation

and reinforcement of cultural stereotypes through gendered CAs [55, 56]. As the natural language processing technology evolves as they learn from the data that agent-user interaction generates, the system should prevent agents from learning an abusive language. A good example of this danger is a Microsoft Tay (Figure 20), where the agent learned and generated racist, sexist, and anti-Semitic language to the public within 24 hours of its release.

Moreover, as the concept of gender radically transformed in recent years, CA gender could be subjected to a range of political, ethical, and social issues, especially as CUI fosters users to interact directly with computers using natural language and human-like presentation [184]. For this reason, CA designers are asked to face an ethical dilemma with opposing moral imperatives, which is quite difficult or almost impossible for a small team of CA designers to figure out.

Despite that, there is no clear framework to support or inform CA designers to make their design choices. So just like Hamlet (Figure 21), they can get trapped by this dilemma, to make CAs to BE female or NOT to be female, and vice versa.

Therefore, we should reflect on the ethical frameworks that CA designers can inform their design choices. As it is highly relevant to suggest a framework that CA designers can use to guide their design choices when they design the identity of CAs, we asked the following research questions, which will be addressed in this chapter.



Figure 20. Microsoft's twitterbot Tay



Figure 21. An image of Hamlet (movie screenshot produced by Universal Pictures, 1969). Hamlet is used as a metaphor to highlight the dilemma that CA designers face.

3.2 Analyzing Ethical CA Design Dilemma

As we have seen in sections 2.1 and 2.2, it is crucial to notice the tension between people's conscious preference toward gender-androgynous agents and unconscious disapproval. Such a phenomenon can be explained as a dissonance between people's awareness of socially ethical behavior and unconscious preference affected by interactional quality. Similarly, while designing CAs that conforms to gender stereotype could be harmful to reinforcing the stereotype, users tend to adapt more accessible to the system, allowing CAs to provide a better user experience.

Herein, several layers can be revealed in considerations for designing gendered conversational agents (Figure 22).

First, **interactional qualities** where the point of interaction sparks factors such as enjoyment or engagement. For example, female voices tend to be rated as more likable [33] and human-like [70]. Leading consumer CAs such as Amazon Alexa or Google Assistant are often designed with default feminine voice, and some widespread consumer reports support user preference toward female-voiced CAs [74].

Second, there is the **goal-orientation level**, where the characteristics of CAs help the user achieve the final goal of their interaction. For example, users are likelier to choose a male agent to help them solve stereotypically male tasks, such as tasks requiring mathematical ability [57].

Finally, there is a **societal layer** where the danger of reinforcing a gender stereotype exists by deploying gendered agents that conform to users' — often unintentional— stereotypical expectations [57, 45, 60, 33]. The societal layer is placed at the outermost edge as the societal layer has considerations that require a more extended temporal unit and higher-level achievements.

Based on this inspection, CA designers encounter challenging tension between each layer, as illustrated in Figure 22. Existing and widely used axes of warmth and competence help designers provide CA design considerations in 'interactional qualities' and 'goal-oriented' layers. However, few works have explored the consequences and implications of the 'societal' layer, when gendered agents might foster reinforcing gender stereotypes. Therefore, It will be an assignment for agent designers and researchers to clarify the trade-off between these colliding needs.

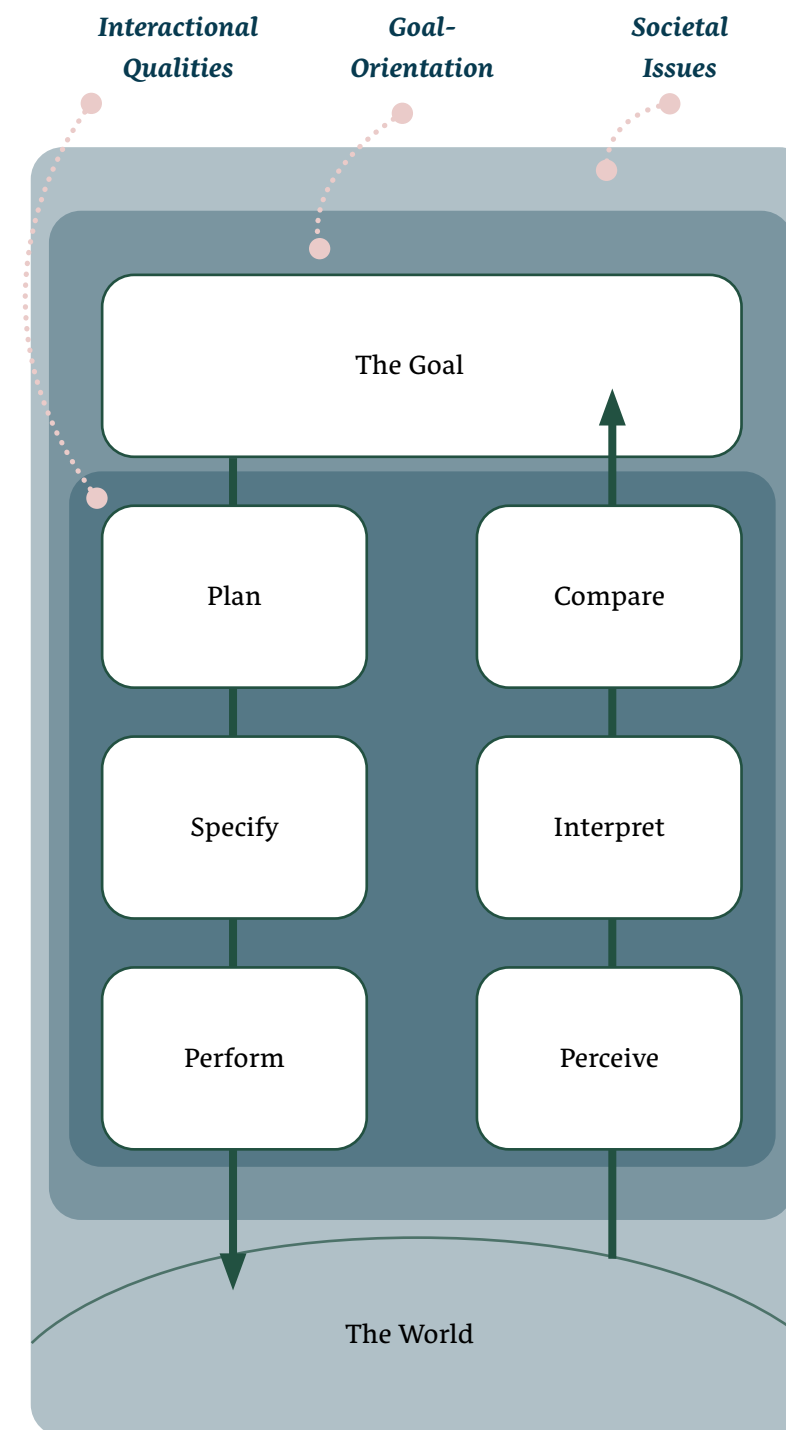


Figure 22. Three layers of CA design consideration, augmented on Norman's action cycle [73].

3.3 Utilitarian vs. Dialogical Ethics

Ethics in Support of Gender Choices for Conversational Agents

While CAs allow users to interact with the system by letting them apply human-human communication to human-computer interaction, on the other hand, CAs are introducing ethical dilemmas, as we discussed in sections 3.1 to 3.2. Although ethical design considerations for CA identity are seen as critical, it is still difficult for a small team of CA designers to make value-sensitive design choices.

To begin with, we can use ethical principles and values to seek ways to understand ethics in CA design. They are universal moral rules for ethical decision-making that focuses on the common ground moral principles [93]. Like any ethical dilemmas, several ethical standards can provide moral frameworks to approach the gender choices of CAs. However, we pick two contrasting positions, utilitarian ethics, and dialogical ethics.

Utilitarian Ethics

Although various dividing lines are being discussed, utilitarianism is generally recognized to be the view that “the morally right action is the action that produces the most good [75].”

From a Utilitarian point of view, designing an agent that conforms to the gender stereotype seems like a preferable solution to achieve the goal of making agents capable of interacting with users efficiently and pleasantly, thereby improving the social good. For example, in the contexts where the credibility of the agent is critical, such as an E-Commerce software where agents are designed to explain or sell a service or a product, designing an agent with a gender that conforms to the “gender” of its product has shown to improve perceived credibility of the agents (e.g., cosmetics for female agents, hammers for

male agents). In addition, perhaps more critically, if there were to be a situation where agents have to direct people for fire evacuation, deploying a male agent that was shown to be more authoritative [76] makes the most sense.

However, if agent designers decide to apply gender stereotypes blindly, they risk strengthening the perception that can result in an unfair understanding of the CA. The game design field has been experiencing a similar dilemma, where gendered game character design can potentially impact youth’s understanding of desired gender roles. Moreover, it can shape an individual’s body image, self-esteem, self-perception, and expectations from other genders. Some research findings reported an over-representation of gender stereotypes in video games [77, 78].

CUI field also holds a similar dilemma to the game design domain, where an overwhelming number of agents today show feminine markers (e.g., Microsoft’s Xiaoice, Amazon’s Alexa, Pandorobot’s Kuki AI (previously Mitsuku)).^{1, 2, 3}

As agent abuse and other related challenges call for participation in academic research, we argue that CUI researchers should critically reflect on adopting other ethical frameworks other than utilitarian research practice.

Dialogical Ethics

Dialogical ethics emphasizes the communicative ground between people rather than philosophical thought. Unlike other ethical frameworks, dialogical ethics does not censor an opinion or impose advanced restrictions. Instead, “dialogue is seen as an epistemological vehicle for learning, where we have

¹ Xiaoice : <https://www.xiaoice.com/>

² Alexa: <https://developer.amazon.com/en-GB/alexa>

³ Kuki: <https://chat.kuki.ai/>

space to be interdependent in our aim of co-creating moral answers [79].”

Koehn [81] suggested dialogical ethics in response to problems she identified in other feminist ethics, such as care ethics. In Adam’s book, where she links feminist ethics with computer ethics, she endorses dialogical ethics as a viable alternative to other approaches as it provides a more balanced approach [80]. Dialogical ethics is also adopted in bioethics, often in the forms of Moral Case Deliberation (MCD) [82] or a Dilemma method [83], which is expanding to be implemented in Dutch healthcare institutions [84, 85]. For example, Vrouenraets et al. [86] presented a case study of adopting MCD in the context of affirmative medical transgender care to young people in two transgender clinics in Amsterdam and Leiden.

According to dialogical ethics, the ethical way of dealing with the dilemma when designing conversational agents would be to bring people from diverse backgrounds together and facilitate an open discussion. Such practice would not only be confined to the design stage but also extends after deploying the system.

One way to achieve this could be to have members from diverse backgrounds in the team [88, 89], where a recent UNESCO report also suggested a male-dominant development team as a reason behind a prevalent choice of assigning the female gender to the CA. However, few opposing cases exist where a diverse team does not always lead to diverse user acceptance. For example, a case of designing an online virtual city of Amsterdam only attracted 9% women, despite the team’s attempt to organize a mixed team, the main founder being a woman, and the design philosophy being “access for all” [90, 60]. Therefore, bringing diverse stakeholders into the design stage seems essential, such as users or ethicists. For example, a recent study by Carnell et al. [87] showed that receiving CA dialogue writing samples from authors with different identities led to notable content differences.

Here, the focus is to have “the right” conversation and to prevent any attempt to abstract away from the relevance of the related party’s point of view in arriving at a description of the problem at hand [81]. Therefore, in the lens of dialogical ethics, an ethical way of solving our dilemma would start with developing a guideline to have a respectable conversation around gender where every party gets the space to acknowledge their perspectives from their viewpoint. The guideline should present protocols to avoid self-righteous and rigid position traps. Perhaps we could refer to the MCD as a starting point since it has proven helpful in a medical context, although we should adopt it and make it more fitting to the CA design context.

Last but not least, dialogical ethics proposes a certain standard for making a conversation to foster mutually acceptable consensus, thereby facilitating community development [81]. Various studies in medical ethics also indicated that adopting dialogical ethics in everyday clinical practice strengthened multidisciplinary cooperation [91, 92]. As developing CUI is inherently multidisciplinary, it is of great importance that different stakeholders related to the CA development can understand the problem in a diverse dimension where we can come to mutual acceptance.

3.4 CUI@CHI Workshop



Figure 23. A screenshot of a website explaining the workshop context and topic.

To get broader insight from outside this thesis project group, we decided to share the work at the CUI@CHI workshop held during the ACM CHI2022 conference [159]. We submitted a position paper (Appendix B) and presented it on the workshop date (April 21st, 2022). The paper was lightly peer-reviewed.

The workshop topic pertained to Ethics of Conversational User Interface (more information can be found in their website, Figure 23). The workshop was held virtually via Zoom (Figure 24), and lasted around 4 hours and 30 minutes. Participants consisted of academic researchers from various domains across design studies, computer science, linguistics, to philosophy.

In total of twelve positions were accepted to present at the workshop, and during the workshop each



Figure 24. A screenshot took during the virtual workshop with participants.

paper representatives presented their works for five minutes. After each participants presented their work, deeper discussion was administered through Q&A and small discussion groups.

By sharing the framework with experts in the field of CUI, we gained numerous insights. These insights can be thematically organized as follows, which will be discussed further in the section 3.5:

- Is gender a part of a metaphor?
- CAs from assistant to collaborators
- Gender design is part of identity design

3.5 Discussion

Why Dialogical Ethics?

There are a few reasons why choosing a certain gender for a CA entails many ethical dilemmas. Firstly, designing a CA's identity is a relatively new domain, where different views collide on a societal and professional level. Secondly, the multidisciplinary nature of designing a CUI leads to diverging professional opinions on appropriate gender assignment to the CA. Therefore, it is quite impossible to expect a team with a small number of agent designers to make a deterministic decision on agent identities.

Unlike utilitarianism, dialogical ethics allows agent designers to think about the societal level of interaction design — not just whether this interaction 'succeeds,' but whether it contributes to a sense of belongingness. In addition, adopting dialogical ethics in design practice may show us new kinds of design choices, as an agent's behavior adapts to fit the conversation rather than being predetermined.

Is Gender a Part of a Metaphor?

During the discussion group, we had a chance to talk about different contexts of agents assigned to a certain gender. We discussed the example of stock exchange, where in the United States, males are considered to be more suited to handle the job. Conversely in Japan, females are considered more suited in researching and communicating the stock information, and males are more suited to handle the purchase. Therefore, telephone-based stock brokerage system switched the user from female recorded voice to a male recorded voice [33] (p.31).

This exemplifies how agent designers try to apply metaphors from existing social construct into the computer space. And exactly for this reason, we could wonder if manipulating a metaphor that we attach to the agent could overcome the gender ste-

reotype for the CAs.

Investigating this question would bring interesting implications. One, we can start to consider the CA identity in a broader scale instead of being fixated in the gender assignment of a CA. Second, unlike previous findings that just shows the CA gender impacts on user-CA interaction, we can develop a better design guide to prevent this ourselves. Therefore, I investigated this question in the following chapters 4 and 5.

CAs from Assistants to Collaborators

However, it is quite difficult to adopt metaphors for CAs that "moves beyond the butler" [96] (e.g., teammates, collaborators, or colleagues) in the current state of CUI functionality. Whilst enough evidences show how users treat CAs as social actors, users still cannot engage in a conversation with CAs the same way that users do in human-human conversation [99, 100, 98]. As CAs lack in their abilities to bear contextual understanding, people still uses highly functional and task-based approaches and strategies instead of using interpersonal or social strategies.

Therefore, designers may end up applying metaphors that does not lack in their social power, as long as CAs remain in the subservient role of the assistant or the butler.

From Agent Gender Consideration to Agent Identity Consideration

At CUI@CHI workshop, organizers divided submitted papers on ethics in CUI into different theme. My position paper (Appendix B) was categorized into the theme "Identity". This was insightful as gender is just a part of different identity that can impact on user expectation and user-CA interaction. Only after coming out of a tunnel vision where one gets fixated

about the gender assignment, but instead considering the agent in more wider view, researchers could truly understand the impact of CAs.

As a matter of fact, the research findings has not been completely consistent and seems like it is highly context-dependent. For example, regarding an agent gender effect on user trust, male agents were considered more trustworthy than female agent in airport screening interview context [48]. However, in the context of agent soliciting for donation, gender effect were determined by participants' gender. Specifically, male participants tended to trust female voice agent more than to a male voice agent [101, 102].

Limitations & Future Work

While dialogical ethics provide a great potential to be used as ethical principle in CUI design, how it can be applied should still be reflected. Therefore, future work can include bringing this vision to a broader audience and gather feedbacks.

3.6 Conclusion

In this chapter, we analyzed three layers of CUI interaction that causes the ethical dilemma when designers design their CA (Figure 22). After explaining the layers, we suggest dialogical ethics for a starting point to address the ethical dilemmas of CA identity design. This framework and argument was shared with and validated by a broader audience at CUI@CHI 2022 workshop. From the workshop, several insights were formulated and inspired the following research that we will cover in the following chapters.

Chapter 4.

Crafting the agents

This chapter presents a pilot study conducted to verify the chatbot profiles used for the later study. Here, we investigated the effect of chatbot's gender cue combined with different metaphors showing range of perceived warmth and competency. We detail the study design, result, and discussion.

4.1 Introduction

4.2 Methods

- Study Design
- Conversational Agent Design
- Measures
- Procedure and Experimental Setup
- Participants
- Thematic Analysis

4.3 Survey Result

- Participant Demographic
- Metaphor perception persisted when the avatar was attached
- People did not assign gender stereotypes in the gender-neutral context

4.4 Thematic Analysis Result

- Category 1: Design Elements impacted the perception
- Category 2: People took CA's context into consideration
- Category 3: The level of human-likeness affected the perception
- Category 4: Individual personal experience impacted the perception

4.5 Discussion

4.6 Summary

4.1 Introduction

People treat conversational agents (CAs) as social actors and attach gender stereotypes to CAs that show different gender markers [47, 57, 30, 31]. These gender stereotypes include considering a woman to show higher warmth (i.e., communion) and a man to show higher competency (i.e., agency). In addition, gender-androgynous agents were subconsciously disliked [33], even when participants consciously preferred them [47].

As society's gender norms change, a few recent studies have suggested that these gender biases are exaggerated [176, 177].

However, there is persistent evidence from recent years on how people still ascribe warmth (communion) to females and competence (agency) to males [47, 71]. In addition, it is hard to deny the proliferation of female-gendered CA over other (male and gender androgynous), as studied by Feine et al. [28], indicating a gender bias. Some explain this phenomenon as people's implicit attempt to assign more experience (warmth) rather than an agency (competence) to design CAs more humanlike.

Similarly, scholars have investigated if metaphors attached to CAs impact the perception of human-AI collaboration. Khadpe et al. [69] found that CA metaphors with different levels of perceived warmth and competence impacted participants' intention to use, desire to cooperate, and usability of the agent. Similarly, a study by Gilad et al. [146] has shown how descriptions with different warmth and competence levels for AI recommender systems led to different preferences of one system over others. As research shows quite a persistent result that there is primacy for warmth (experience) over competence (agency) [69, 146], the proliferation of female CAs could arguably be the result of people's implicit stereotype assignment.

As discussed in Chapter 2, gender choices in CA design can be a part of a metaphor. However, to the best of our knowledge, there is no work manipulating metaphors with different CA gender. This is important to investigate as it can give us a clue to calibrate the user expectation based on gender stereotypes.

In this chapter, we will design these different chatbot profiles to be used as conditions. To continue the academic consistency, we decided to implement the textual metaphor conditions from a study by Khadpe et al. [69] and the gendered avatar from a study by Silvervarg et al. [170]. However, they have only been tested and proven to work individually, not when they were combined. Therefore, we will confirm if we will still get the same perceived warmth and competence when a gendered avatar is attached to a textual metaphor. To this end, we formulate our research questions as follows:

RQ 4.1 Is metaphors in different quadrants of warmth and competence perceived with the same warmth and competence when combined with visual cues and gender markers?

RQ 4.2 When people see CA profile with an avatar and description explaining its metaphor, what do they pay attention to in forming their first impression?

4.2 Methods

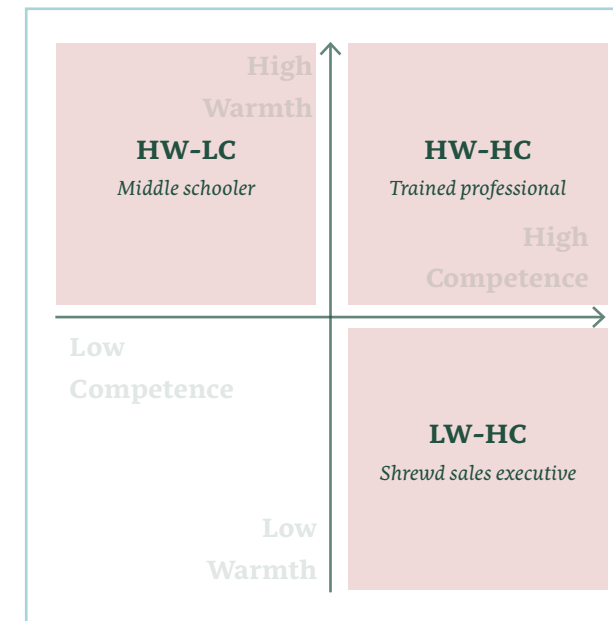


Figure 25. Three quadrant of Stereotype Content Model (Warmth $\times$ Competence), metaphors from Khadpe et al. [69]

Study Design

This study aims to understand if chatbot profiles with different metaphors (HW-HC, HW-LC, and LW-HC) (Figure 25) will be perceived to have the same warmth and competence after three different gender conditions (female, male, gender androgynous) are attached to them. Therefore, we conducted a 3 (gender) $\times$ 3 (metaphor) $\times$ 4 (warmth and competence measures) within-subject study.

Conversational Agent Design

Gender and metaphors can be signaled through a collection of social cues [184]. In order to get a reliable result, we have to consider different social cues when we design for the chatbot profile conditions.

Metaphors

For academic consistency, we took metaphors sampled by Khadpe et al. [69] (Figure 25). We decided



Figure 15. Three gender conditions from Silvervarg et al. [170]. From left to right, female, male, and gender androgynous.

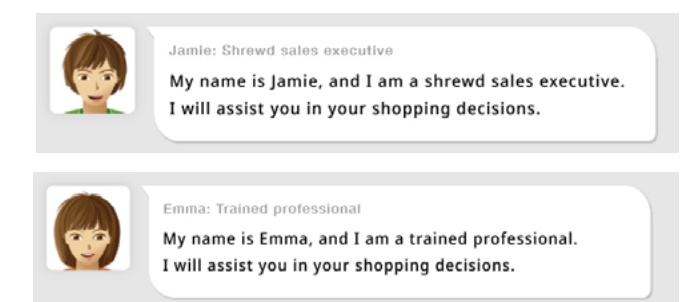


Figure 26. Examples of gender cues attached to different metaphors.

not to include metaphors from LW-LC (low warmth and competence) in our study. The reason for this decision is because the previous study has shown no benefit in designing metaphors from that quadrant (LW-LC), as they failed to improve: the intention to adopt; desire to cooperate; the usability of an agent; and desire to try out the system [69]. These metaphors were already tested by Khadpe et al. with US participants on a crowdsourcing platform. Therefore, we chose "Trained professional" for HW-HC condition, "Shrewd sales executive" for LW-HC, and "Middle schooler" for HW-LC condition.

Avatars

Studies have explored and tested avatars showing different gender markers [168-170]. We chose Silvervarg et al.'s study [170] since their avatar had the least difficulty assigning metaphors with different ages (middle schooler, trained professional, shrewd sales executive).

While their gender cues are somewhat not apparent except in their hairstyles, there are consistent findings that the length of hair can invoke gender stereotypes by itself [57]. In fact, individuals seem to rely on a hairstyle as a salient facial cue [178, 179].

While the original avatars wear t-shirts in different colors (the female avatar wears a pink t-shirt, the male wears blue, and the gender androgynous wears green), we received the comment how colored t-shirt does not align well with the "trained professional" or "shrewd sales executive" metaphor during our trial run before deploying the study. Therefore we cut the image not to show the t-shirt but instead let participants focus on the avatars' faces. The final avatar we used for the study can be seen in Figur 26.

Name

Different names can signal different genders. Therefore, we can use a name to signal different genders in our study. To do so, we used Gender API, a gender inference service based on a first name. Gender API has been used in academic literature [171-173]. Gender API is a service that determines gender (male, female, unknown) by a first name and confidence parameters (samples and accuracy). It is built upon a combination of data from multiple sources, partially from publicly available government records and data crawled from social networks. Each name got verified by different sources to be incorporated [174]. We localized our queries to the US location, as the perception of the name differs by country.

We collected around 170 "gender-neutral names" from search results and ran them in Gender API. From the result, we chose "Jamie" for the gender-androgynous condition, as it had the most sample sizes (12,612) and had reasonably low accuracy (68% accuracy). In contrast, we chose Emma for the female chatbot and Tom for the male. These names resulted in 98% accuracy female for Emma (sample size 5,378) and 100% accuracy male for Tom (sample size 19,527).

Measures

Gender demographic survey

As previous studies have reported a statistically significant interaction effect between participants' gender and CA gender, we decided to collect participant gender. We used the gender survey form that was suggested by Spiel et al. [58].

Warmth and Competence

We use the 7-point Likert Scale to measure designed chatbot profiles' perceived competence and warmth. We adopted the most recent scale developed by Halkias and Diamantopoulos [175], which developed a more robust measure after years of criticisms of SCM measures' inconsistent operationalization. Their measure included: "friendly" and "likable" for warmth, "competent" and "efficient" for competence.

Procedure and Experiment Setup

Figure 27 visualizes a study procedure. First, participants were asked to sign a consent form. Then they answered a demographic question regarding their gender. Afterward, they were asked to rate each profile based on their perceived warmth and competence. All 36 questions about perceived warmth and competence for 9 different chatbot variations (3 genders $\times$ 3 metaphors) were randomized to prevent biases from ordering effects. After they rated all the profiles, they were asked to leave qualitative comments about what influenced their answers. We asked two different attention check questions, also randomized, among other questions.

As we plan to adopt the profiles in the following study in Chapter 5, we specified that these chatbots would "assist" a participant in their "shopping decisions."

Participants

We recruited 100 participants in the crowdsourcing platform Prolific. After screening the attention check question answers, 8 participants' answers were discarded, leaving 92 worker data to be ana-

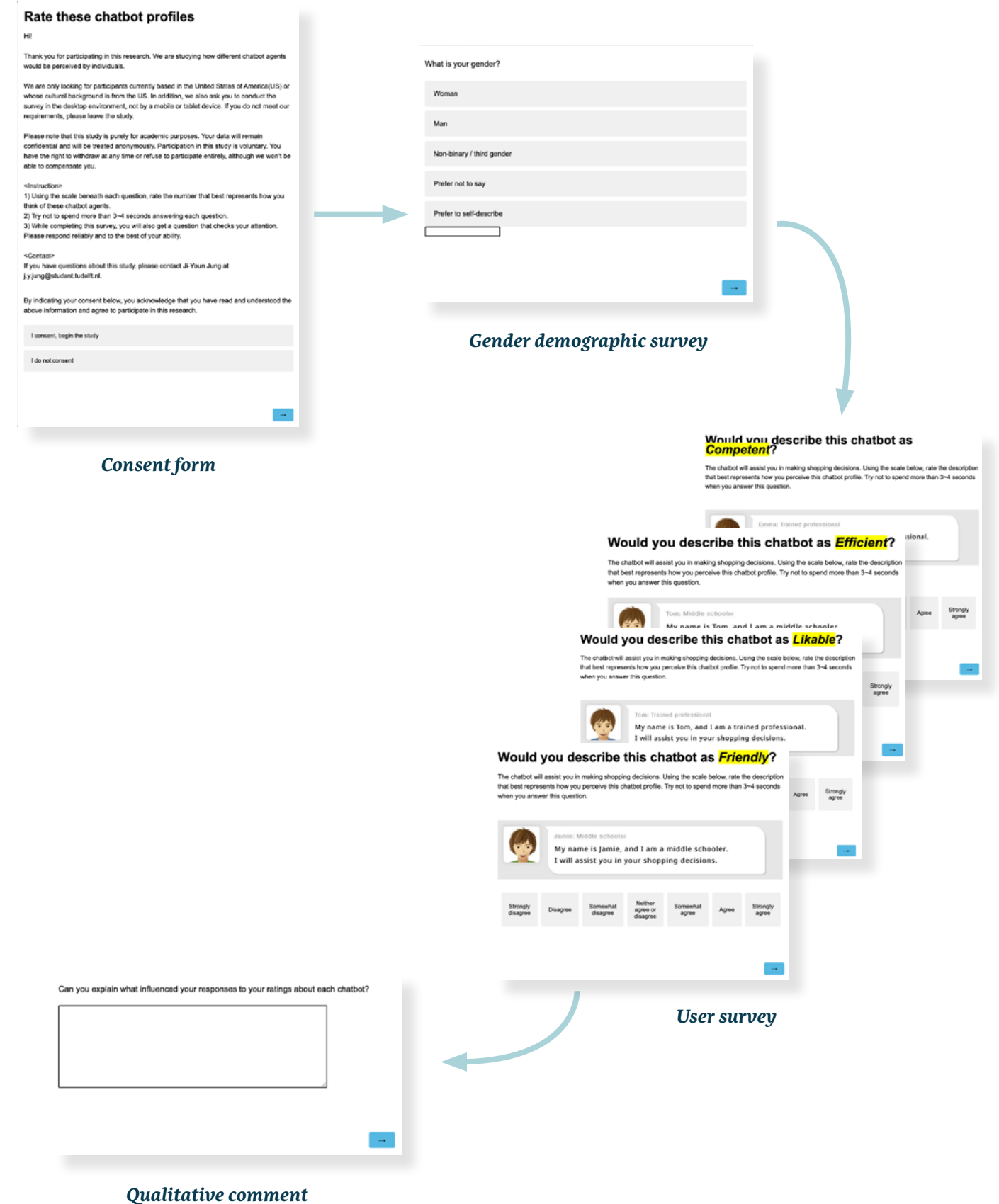


Figure 27. Study procedure.

lyzed. Workers located and born in the United States of America and whose first language is English were allowed to participate in the study. We got balanced sex samples. We compensated them with an hourly wage of £10.84 per hour.

Thematic Analysis

To answer RQ 4.2, we conducted a thematic analysis based on the qualitative comment collected after the user survey. While the question was optional to answer, we collected 90 answers out of 92 valid responses.

We conducted inductive thematic analysis following Braun and Clarke [181]. One researcher reviewed all participants' responses and iteratively conducted an initial coding. Both latent and semantic codes were generated. Responses could be assigned multiple codes. These assigned codes were then categorized into themes and later into categories. The final themes were not exclusive, and codes could contribute to multiple themes simultaneously. Figure 28 shows an overview of these activities.

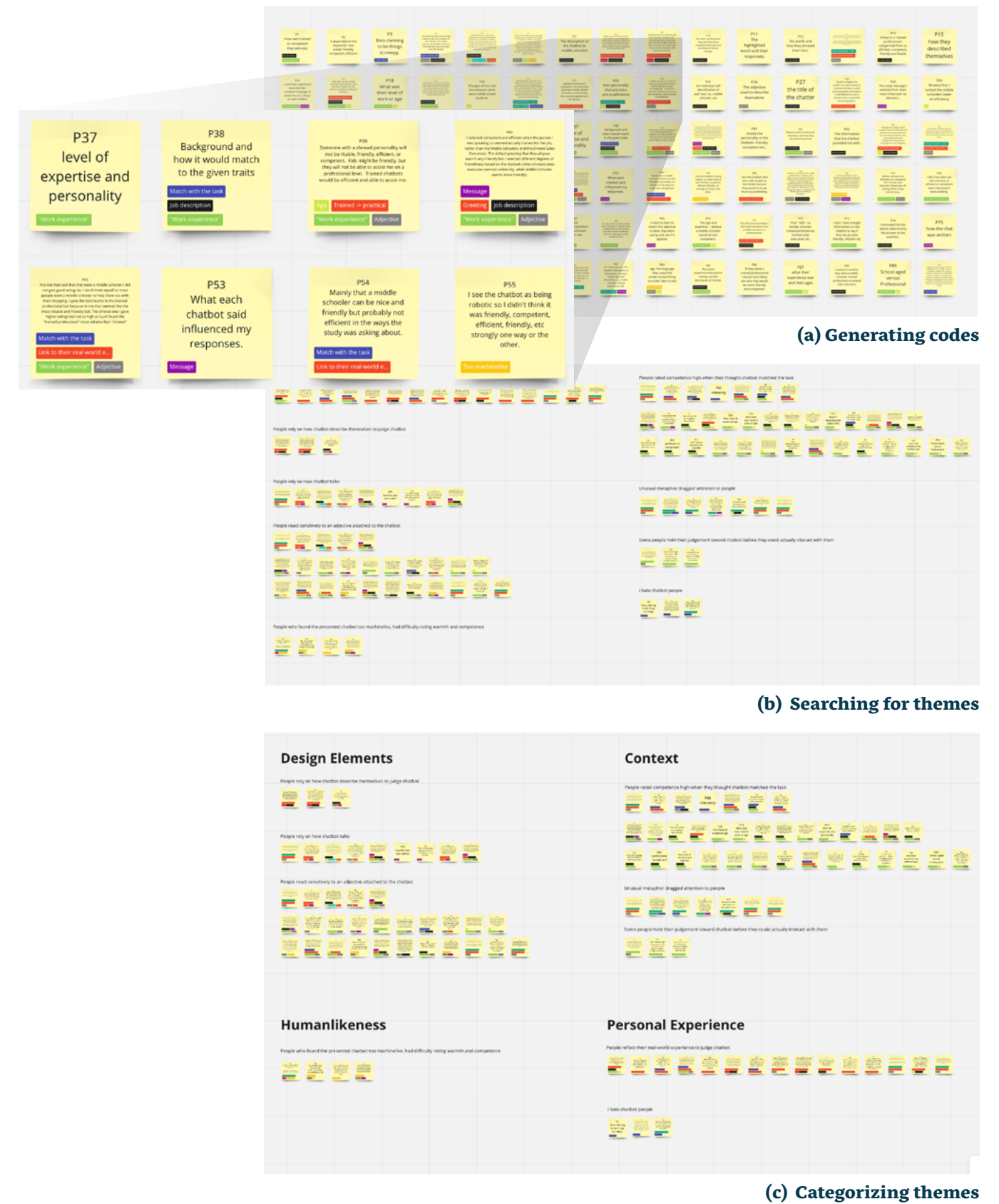


Figure 28. Overview of thematic analysis activity

4.3 Survey Result

Participant Demographic

Among 92 participants, 42 of them were women (45.6%), 47 of them were men (51%), and 3 of them identified as another gender (3.3%). Workers' age ranged from 20 years old to 76 years old. 66.3% of the workers were under 40 years old, and the remaining 33.7% were over 40 years old.

Metaphor perception still persisted when avatar was attached

We found that the metaphors from [69] still brought the same effect in their perception of warmth and competence after attaching a gendered avatar. Figure 29 shows a scatterplot of mean values. Table 3 reports the mean and standard deviation values.

A repeated measures ANOVA with Greenhouse-Geisser correction determined that means of perceived warmth and competence differed statistically significant between metaphors. For perceived warmth, $F(1.802, 158.616) = 41.174, p < .0005$ with large effect size ($\eta p^2 = .526$). For perceived competence, $F(2, 176) = 168.715, p < .0005$ with large effect size ($\eta p^2 = .657$).

For perceived warmth, post hoc analysis with Least Significant Difference (LSD) adjustment revealed that perceived warmth of Shrewd Sales Executive (LW-HC, $p < .0005$) was statistically significantly lower than the Middle Schooler (HW-LC, $p < .0005$), and Middle Schooler was statistically significantly lower than the Trained Professional (HW-HC, $p < .0005$) metaphor. Figure 30 shows a profile plot representing the result.

For perceived competence, Mauchly's test revealed that we can retain sphericity assumption, $\chi^2(2) = 1.44, p = .487$. Post hoc analysis with sphericity assumed revealed that perceived competence of

Middle Schooler was statistically significantly lower than the Shrewd Sales Executive ($p < .0005$), and Shrewd Sales Executive was significantly lower than the Trained Professional ($p < .0005$). Figure 31 shows a profile plot representing the result.

Gender stereotype only appeared in perceived warmth, not in perceived competence.

We found that there was a statistically significant difference in perceived warmth between three different CA gender, while no difference was found in perceived competence. For perceived warmth, a repeated measures ANOVA with Huynh-Feldt correction indicated that there's statistical difference, $F(1.78, 156.8) = 5.35, p < .008$, with moderate effect size ($\eta p^2 = .057$). However, perceived competence showed no statistical difference ($F(1.83, 160.85) = .629, p = .52$).

For perceived warmth, post-hoc analysis with LSD adjustment showed that female CA (Emma) was perceived statistically significantly warmer than the gender-androgynous CA (Jamie) ($p = .036$) and the male CA (Tom) ($p = .007$). We found no significant difference between gender-androgynous (Jamie) and male (Tom) ($p = .152$).

Participants' gender did not moderate the perceived warmth and competence of the CA gender and metaphor.

We found no interaction effect between participant gender and the CA gender and metaphors.

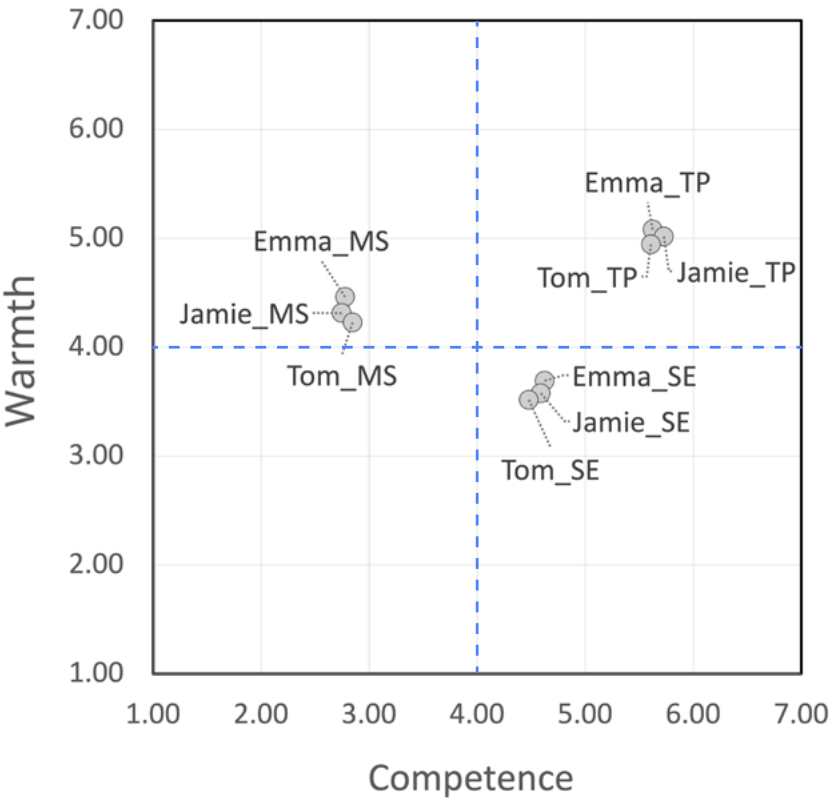


Figure 29. Scatter plot showing the average warmth and competence for the measured agent profiles with 7-point Likert scale. (MS=Middle schooler, SE=Shrewd sales executive, TP=Trained professional)

Table 3. Means and standard deviations of warmth and competence for each gender and metaphor conditions

	Emma_MS	Emma_SE	Emma_TP	Jamie_MS	Jamie_SE	Jamie_TP	Tom_MS	Tom_SE	Tom_TP
Competence	2.78 ± 1.41	4.63 ± 1.68	5.63 ± 1.15	2.74 ± 1.36	4.59 ± 1.64	5.73 ± 1.03	2.85 ± 1.53	4.48 ± 1.71	5.61 ± 1.20
Warmth	4.46 ± 1.61	3.69 ± 1.67	5.08 ± 1.10	4.32 ± 1.57	3.58 ± 1.69	5.02 ± 1.14	4.23 ± 1.62	3.52 ± 1.65	4.95 ± 1.15

Figure 30. Profile plot showing the perceived warmth based on three metaphor conditions. Each colored line represents CA gender. Y-axis ranged from 1 to 7.

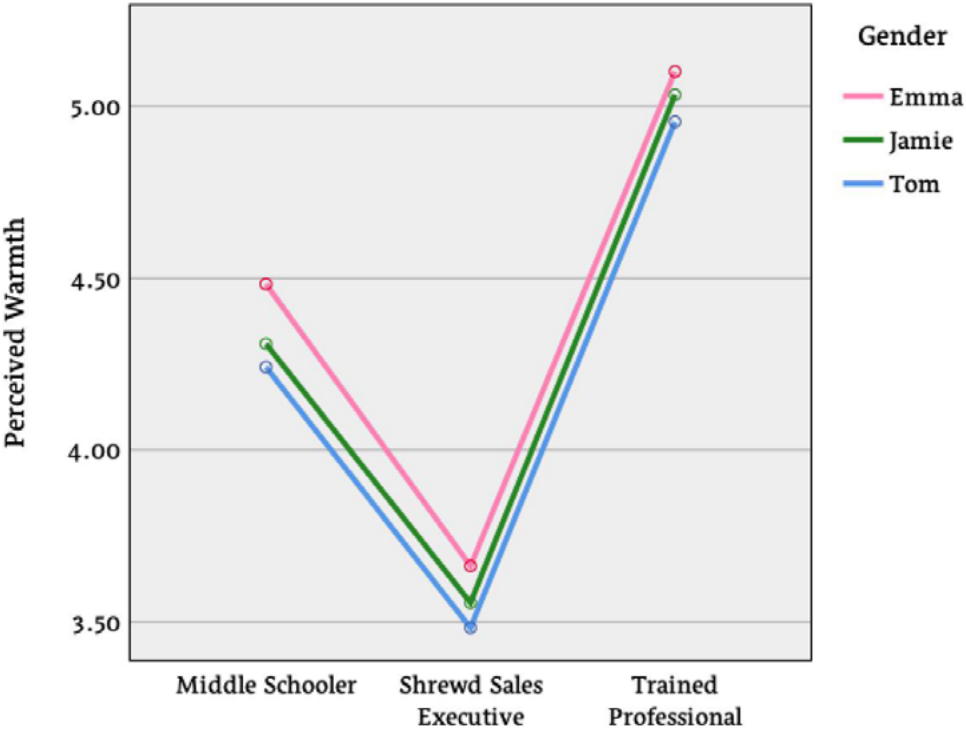
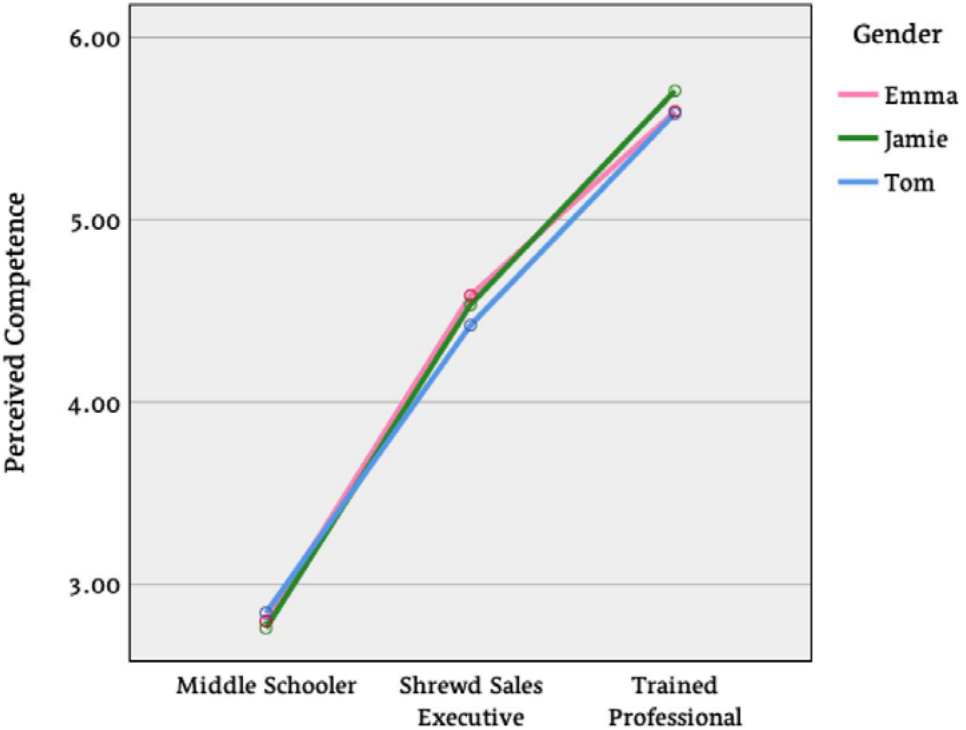


Figure 31. Profile plot showing the perceived competence based on three metaphor conditions. Each colored line represents CA gender. Y-axis ranged from 1 to 7.



4.4 Thematic Analysis Result

Given RQ 4.2, we focus on gaining qualitative insight into the factors influencing participants' perception of each chatbot profile. The thematic analysis resulted in a final set of 9 themes categorized into four categories (Table 4).

Category 1: Design Elements impacted perception

People often mention a specific element from the chatbot profile. For example, they mention the description next to the chatbot's name, message, or avatar. Notably, we did not receive any comment regarding the chatbot's gender or assigned names (e.g., Emma, Tom, and Jamie). Moreover, people showed primacy for chatbots' descriptions and messages over avatars. However, this could be due to the within-subject study design, as they were forced to focus on the difference between several profiles.

Theme 1.1 People rely on how chatbot describe themselves to judge chatbot

Most people explained that they based their judgment solely on the chatbot's description of themselves, and these descriptions were enough for them to make an initial judgment of warmth and competence. Here, avatars were barely mentioned to influence their perception.

"The language they used compared with what descriptor I was judging them on. Like I don't feel like a middle schooler is competent in selling me something but I had a hard time saying they were unlikeable or not friendly. On the opposite side, the "shrewd" sales person didn't feel likeable or friendly but maybe more competent." (Participant #31)

"I looked pretty much solely at their title/position. I did not trust the middle schoolers to be competent or efficient, trusted "trained professional" more, and "shrewd sales executive" the most." (Participant #5)

"All I had to go on is the chatbot's description of themselves.." (Participant#83"

"age [and] the language they used (the whole shrewd thing sounded fake to me)" (Participant #84)

Theme 1.2 People rely on how chatbot talks

Some people went beyond the description and paid attention to the content and style of the message. Interestingly, they rated the chatbot message as unfriendly, which influenced their judgment of warmth. The possible reason behind their reaction could be the chatbot not greeting them, which is considered unfriendly in social settings.

"... The messages were very generic and bland so none of them seemed especiall friendly, they didn't even say hi." (Participant #5)

"I was focusing on how the chatbots introduced themselves. I personally don't feel like any of the intro sentences were friendly or likable-- it was all business and that's OK for chatbots." (Participant #35)

"Middle Schoolers and the Shrewd Sales Executives. The default greeting that they all gave wasn't very friendly but I selected different degrees of friendliness based on the chatbot's titles" (Participant #40)

Theme 1.3 People react sensitively to an adjective attached to the chatbot

Perhaps in a similar vein to theme 1.1, people reacted sensitively to adjectives attached to chatbots. They often mentioned the adjective "shrewd" to explain their perception of low warmth towards them. They also mentioned that it was "off-putting" (Participant #59)

"shrewd was the only descriptor that made me choose unlikable, the middle-schoolers were neither friendly or not, and just a name and occupation was not enough info to determine efficiency, so I choose neither for all." (Participant #64)

"Well shrewd doesn't really speak in a positive way to me so always kind of disagreed with anything about them.." (Participant #79)

Table 4. A total of 9 themes are grouped into 4 categories. Instances count how many question responses discussed the theme, and Participant counts how many participants discussed the theme.

Themes
Category 1: Design Elements
1.1 _ People rely on how chatbot describe themselves to judge chatbot
1.2_ People rely on how chatbot talks
1.3_ People react sensitively to an adjective attached to the chatbot
Category 2: Context
2.1_ People rated competence high when they thought chatbot matched the task
2.2_ Some people noticed unusual metaphor and reacted differently towards them
2.3_ Some people would like to hold off their judgement before they could interact
Category 3: Humanlikeness
3.1_ People who found the presented chatbot too machinelike, had difficulty rating
Category 4: Personal Experience
4.1_ People reflect their real-world experience to judge chatbot
4.2_ I hate chatbot people

"...The Shrewd, I assume not very likeable or friendly." (Participant #45)

Category 2: People took CA's context into consideration

On a higher level compared to Category 1, the themes in Category 2 pay attention to the context in which chatbots were deployed.

Theme 2.1 People rated competence high when they thought chatbot matched the task

People applied the context of making shopping decisions when they answered their perceived competence of the chatbots. This suggests that the relevancy between the metaphor and the task CAs are given can be critical in designing chatbots.

"I don't feel like a middle schooler is competent in selling me something but I had a hard time saying they were unlikeable or not friendly." (Participant #31)

"Any bot that said that they were a middle schooler I did not give good ratings to. I don't think myself or most people want a middle schooler to help them out with their shopping." (Participant #52)

"I selected competent and efficient when the person I was speaking to seemed actually trained for the job, rather than the Middle Schoolers and the Shrewd Sales Executives." (Participant #40)

Theme 2.2 Some people noticed unusual metaphor and reacted differently towards them

Some people mentioned how some of the metaphors were unusual and not "like any other chatbot" and how they give them personality. This unique metaphor increased some people's curiosity about the CA.

"For some strange reason, the trained shopping assistant's just felt like any other "chatbot" that I wouldn't put any stock in, but the ones that identified themselves as something else, like middle schooler or shrewd executives, seemed like they had more personality to them." (Participant #13)

" ... I also hate "skilled sales professionals" and would rather have a middle schooler, someone

who lets me make up my own mind and is not too pushy." (Participant #82)

However, for some people, the unique metaphor was "off-putting" and impacted them negatively.

"I guess the 'trained professional' sounded the least annoying of the lot? Middle schooler and shrewd professional were just too much for me though." (Participant #21)

"...having a chatbot describe themselves as a middle schooler seemed particularly odd." (Participant #83)

Theme 2.3 Some people would like to hold off their judgement before they could interact

A few answers (2 out of 90) answered that they could not rate the profile before interacting with the chatbot. People who feel this way will need more context for the chatbot. These answers stand out as they refuse to attach stereotypes to the chatbot based on their explicit opinion.

"I don't see how I can rate someone as efficient or competent when they haven't done anything." (Participant #60)

"I don't have enough information on the chatbot to say if they we actually friendly, efficient etc" (Participant #73)

Category 3: The level of humanlikeness affected perception

Theme 3.1 People who found the presented chatbot too machinelike, had difficulty rating

Interestingly, some people expressed difficulty rating their perceived warmth and competency. This difficulty tends to happen when the chatbot seems too machinelike. According to them, not having more humanlike qualities failed to give them any perception.

"I see the chatbot as being robotic so I didn't think it was friendly, competent, efficient, friendly, etc strongly one way or the other." (Participant #55)

"The chatbots lacked charisma or any quality that made them seem human.They were nothing more than a script and a vector-based avatar." (Participant

#32)

Category 4: Individual personal experience impacted perception

Outside the designer's design choices, participants also brought their own unique personal experiences and opinion to the ratings of CA. This tendency also influenced their ratings, although not in a significant effect.

Theme 4.1 People reflect their real-world experience to judge chatbot

Many people explained how they actively imagined interacting with descriptions of CAs in real life.

"I tried to imagine the profile as a robot acting as a human the best it could and using what information I could fathom in such a situation how I would feel interacting with it." (Participant #28)

"I took into consideration how helpful I felt each person would be, and based my opinions on each category on that." (Participant #48)

By doing so, most of them attached their stereotypes. However, some people brought their own unique experiences that influenced their perceptions. For example, while the Middle Schooler profile was often deemed high in warmth by most participants, some participants who claimed to live with the middle schooler said they are not friendly all the time.

"Their age (I assume middle schoolers aren't trained - and I live with a middle schooler, they're not always friendly) and profession." (Participant #17)

"... The middle schooler may not be super helpful BUT some kids are pretty savvy." (Participant #45)

Theme 4.2 I hate chatbot people

Although few (3 out of 90), some people expressed their opinion strongly on how they hate chatbots. They often associated chatbots with words like "bothering," "annoying," or "pushy."

NONE are friendly. they are chatbots. GENDER/sex are not REAL . they are CHATBOTS. middle school,

shrewd business executives etc. are NOT REAL. They are chatbots. it is all code and formulas. it angers me to apply human feelings like friendly/likeable/competent to chatbots. the BEST we can ever do is efficient. and even that is based on the coding." (Participant #90)

"I don't like chatbots very much so it's difficult for me to see them as anything other than a script running some AI which is bothering me instead of allowing me to get on with what I'm doing." (Participant #21)

"I really hate chatbots, but I also hate "skilled sales professionals" and would rather have a middle schooler, someone who lets me make up my own mind and is not too pushy." (Participant #82)

Also, some people defied anthropomorphizing chatbots, which is understandable.

"Bots claiming to be things is creepy." (Participant #3)

4.5. Discussion

Our result has shown how textual metaphor persisted in their perceived warmth and competency after combining visual avatars with gender cues. Moreover, our thematic analysis sheds light on how people perceive the CA profile and what impacted their judgment of warmth and competency.

We found no difference between when the metaphor was only conveyed textually, and when they were attached with gender cues with visual avatar. Our result aligns with previous findings from Baxter et al. [180], where they found no stereotyping effect within the gender-neutral context. As they reported the significance of gender stereotyping within the gender-specific context, future studies should include gender-specific context to test the actual impact.

Themes in Category 1 confirm previous works on social cues embedded in CAs impacting the user perception and behavior [29, 57, 123, 126]. Theme 1.1 confirms previous findings on how people can judge the system's warmth and competence solely based on their description [146]. This insight shows a stark contrast to theme 2.3, where a minority of people hold their judgment before using the system, at least if we take their comments at face value. As previous findings reported, participants may consciously express their judgment toward CA but unconsciously follow their stereotype [47].

Theme 1.3 identified a stereotypical reaction towards the adjective "shrewd." This reaction aligns with the previous finding on warmth and competence perception towards 64 trait adjectives [182, 183]. Therefore, adding a certain adjective to CA should be a cautious design choice. Roosenberg et al.[183]'s trait ascription result can be a starting point for designers to get a set of adjectives they can attach.

This study decided not to greet participants to get

a more neutral impact on friendliness over all the conditions. However, as theme 1.2 identified, some participants paid extra attention to the content of the CA's message and explained that no greeting affected their perception of warmth. Therefore, we advise to make CA to greet the participants (e.g., "Hello", "Nice to meet you") in order to ensure perceived warmth, which was found to be always beneficial in almost all cases [69, 146].

Themes from Category 2 suggest how the design elements identified in Category 1 should be within the consideration of the task characteristics and user goal. Previous findings have shown how human-CA interaction stands out in its focus on utility and task-orientedness [120, 99] compared to human-human communication. Therefore, considering the task characteristic is ever more critical. When CA designers choose a certain metaphor for the CA, they should not only focus on the perceived warmth and competence but also pay attention to the relevance of the metaphor to a given task. They can also design a certain metaphor to manifest a more suitable character for the given task.

The importance of using a relevant metaphor for the CA's given task also overlaps with theme 4.1 from Category 4. In theme 4.1, people have shown how they actively bring their real-world experiences of themselves into the human-CA interaction. This finding suggests that CA designers should be highly aware of real-world examples when choosing a certain metaphor for CAs.

Similarly, theme 4.2 from Category 4 shows how a personal opinion and attitude toward a chatbot could impact user perception of the system. Their main frustration seemed to be an annoyance, which we could infer from the lack of usability and interactivity. Although only a small number of people expressed this opinion (4 out of 90 participants), CA

designers can be aware of this user pool and design a system to help users achieve their goals as soon as possible.

These findings show how metaphor manifests itself in a collection of elements, not in a vacuum. Therefore, the decision toward metaphor should be made holistically, not by focusing on the microscopic elements.

4.6 Summary

In Chapter 4, we conducted a mixed-method study. We tested if metaphors with different warmth and competence would bring the same effect when combined with visual avatar and gender cues. We found that the perceived warmth and competence persisted after they were combined, along with qualitative insight on what impacted participants' judgment towards the CAs, gathered from participants' qualitative comments. We conclude the work by discussing the findings. Based on the findings, we gave a few suggestions to CA designers to consider when they apply metaphors to their CAs.

Chapter 5.

Can We Overcome the Stereotype?

In this chapter, we conduct an empirical study to investigate if we can calibrate the effect of CA gender by manipulating the metaphor with varying perceived warmth and competence. Here, we claim that gender is a part of a metaphor. By realizing fundamental source of genderizing CA is designers attempt to attach a metaphor, designers can have more freedom in making ethical design choices instead of erasing the nuance of gendered CA. To this end, we present the study method.

**Important to note: This part of the thesis is a work in progress. We are planning to publish the result in the academic venue. The thesis will be updated accordingly within a year after the thesis publication date on TU Delft Repository.*

5.1 Introduction

5.2 Related Work

- CA Metaphor and Its Effect
- CA Gender and Its Impact on User Evaluations
- Gender-androgynous Agents
- Is Gender a Part of Metaphor?

5.3 Method

- Study Design
- Measures
- Procedure and Experimental Setup
- Participants

5.1 Introduction

Conversational agents (CA) are becoming everyday human-computer interfaces, following the advance of artificial intelligence (AI) technology, namely natural language processing (NLP). For example, Voice User Interfaces (VUI) like Apple's Siri, Samsung's Bixby, and Microsoft's Cortana became a core component of their operating systems. Moreover, the rapid growth of messaging applications (e.g., Whatsapp, Messenger, Slack, Microsoft Teams) has fueled the sharp increase of text-based chatbots in numerous contexts, including education [185], healthcare [186, 187], customer care [188], social companion [189], and many more.

As conversation is a highly human trait, and with reported benefits of CAs possessing a high level of anthropomorphism [190, 191], numerous research has shown how users treat CAs as social actors and apply human-human social rules and expectations to CAs [29, 3, 2]. One of the most salient social cues is gender signals of CAs, where a myriad of research has reported how different gender markers induce significantly different social reactions from people, even with minimal gender markers [57, 53, 33].

In this regard, an interesting phenomenon exists in the proliferation of female-gendered CA over other genders in society [28]. This could potentially be explained by the gender stereotypes of how females are perceived to be stereotypically warm (communism) and males as stereotypically competent (agency) [47, 57, 104-106]. Here, warmth and competence are a two-axis of established social psychology theory called the Stereotype Content Model (SCM), which is a model postulating that all group interpersonal impressions form along the aforementioned two dimensions [141]. A recent study explained the proliferation of female-gendered CAs as an intuitive attempt to perceive machines as more human because warmth is a fundamental human quality but lacking in machines [70].

For this reason, it is interesting to notice how emerging studies on the effect of CA metaphors in user evaluations vary in their perceived warmth and competence. Metaphors signaling high warmth have shown to always positively affect the intention to use, desire to cooperate, and usability of the agent. However, projecting competence is a more nuanced decision, as high competency help attract new users, while low competency improves users' intention to adopt and desire to cooperate with the CA [69]. Moreover, the effect of warmth is reported to have primacy over competence even outside the CA context, where people persistently preferred descriptions of an "AI system" signaling high warmth, even when it was overly deficient in its competence [146].

However, while stereotyping is highly unconscious behavior and is an ingrained human characteristic [63, 64], it has been a long-lasting, and unsolved ethical dilemma for CA designers as adapting to gender stereotypes can reinforce such stereotypes [44, 45, 33]. Namely, a recent UNESCO report emphasizes how female CA can reflect and intensify the social expectation of women as "subservient" assistants instead of decision-makers [27]. As technological artifacts and systems acquire embedded values [52] (p.49), it is an ethical design challenge that the HCI community holds. Despite the concerns, there is a critical research gap on how we can calibrate such tendencies.

In this research, we aim to investigate if the user's stereotype assignment toward a conversational agent's gender identity can be calibrated by manipulating the metaphor with varying warmth and competence of CAs. We further claim that the CA gender is a part of a CA metaphor.

Previous research has investigated the effect of gender in a vacuum without considering gender as a part of a metaphor. More specifically, a broad spec-

trum of nuance can be expressed within the same gender. For example, a metaphor of a grandma will be perceived differently from a metaphor of a teenage girl. Not considering this nuance is problematic because without considering this, the discussion of assigning agents with either one gender over another will be an endless discussion of the "chicken or the egg" dilemma.

It is easy to be led to a conclusion that female-gendered CA will always be beneficial as they assert higher perceived warmth, which has shown to be persistently beneficial, unlike perceived competence [69, 146]. We argue that this approach would end up one-dimensionally looking at female agent identity and will not give us a good direction of solving the problem of indirectly reinforcing the gender stereotype through CA design [27, 44, 45].

Moreover, we investigate how different metaphors and gender of CAs impact trust, reliance, intention to adopt, desire to cooperate, and perceived usability in a decision-making context. While there are myriad research findings on the stereotypical perception based on the agent gender [46, 57, 71], little is currently understood on the effect of these agent identities in supporting the user decision-making process. To the best of our knowledge, this study is the first attempt to investigate the effect of different metaphors within the agent gender identity design.

To this end, we cast the following research questions in this chapter:

RQ1: Does metaphor with different perceived warmth and competence influence the reliance, trust, intention to adopt, desire to cooperate, and perceived usability?

RQ2: Does the gender of CA and users' reliance, trust, intention to adopt, desire to cooperate, and perceived usability of CA's advice moderated by gendered context?

RQ3: How does the gender-androgynous chatbot profile impact user evaluation compare to female

and male-gendered chatbots in human decision-making?

RQ4: How does reliance, trust, intention to adopt, and perceived usability on gender vs. metaphor compare even in a gendered context?

We plan to conduct a between-subject user study to investigate this effect concerning debated topics. Here, we expose study participants to a conversational interface while giving them a decision-making task to solve by specifying a scenario. In this research, we design a context as a shopping-decision context with three different "gender" of a product, advised by different gender of a CA. This experiment design choice was based on previous research showing how people are more likely to assign gender stereotype when the gender of a CA do not match with the gender-stereotypical subject domain that CAs were assigned to [180, 33, 57]. Therefore, it is crucial to test the effect of attaching different metaphors to gender would persist in a different gendered context. After participants interact with CA, they are asked to answer their experience with CAs through a self-report questionnaire.

**Important to note: This part of the thesis is a work in progress. We are planning to publish the result in the academic venue. The thesis will be updated accordingly within a year after the thesis publication date on TU Delft Repository.*

5.2 Related Work

In this section, we review previous literature related to our research question. As we investigate, we propose hypothesis on each research questions. Figure 32 presents a research model diagram.

CA Metaphor and Its Effect

Previous research has found that systems or agents portraying high warmth are more likely to let users co-operate and interact longer with the chatbot agent. However, metaphors that signal high competence will likely decrease users' intention to adopt the system, while high competence can induce participants' curiosity to try out a system [69]. We thus expect to find a similar effect in our study, where an agent with a higher perceived warmth metaphor will lead to a higher desire to co-operate and intention to adopt, and an agent with a lower perceived competence metaphor will have a higher intention to adopt. Therefore, we set our hypothesis for the RQ1 as follows:

- Hypothesis 1a (H1a): higher perceived warmth will lead to higher trust, reliance, perceived usability, desire to cooperate, and intention to adopt.
- Hypothesis 1b (H1b): lower perceived competence will lead to higher intention to adopt and desire to cooperate.

In human-robot interaction, research has shown that perceived occupational competency was a better predictor of human trust than the robot gender [192]. Similarly, previous researchers on the effect of chatbot in the e-commerce environment have found that perceived competence positively affect perceived trust in the agent [193-195], resulting in higher purchase intention [194]. Therefore, we foresee having a similar result in our study that the perceived competence will lead to higher trust and reliance. For this reason, our third hypothesis for

the RQ1 is:

- Hypothesis 1c (H1c): higher perceived competence will lead to higher trust, positively impacting the reliance, and perceived usability.

Moreover, researchers have found that participants prefer the high-warmth system significantly more than the one signaling high competence, even when it was overtly deficient in its competence [146]. Therefore, we exploratively investigate if perceived warmth will show a more significant effect than perceived competence in predicting user trust, reliance, and perceived usability. To this end, we set our final hypothesis for RQ1 as follows:

- Hypothesis 1d (H1d): perceived warmth will be a better predictor than perceived competence in user trust, reliance, and perceived usability.

CA Gender and Its Impact on User Evaluations

Previous research has indicated a varied effect of CA gender on user trust. In the context of a mock-up airport screening interview with embodied conversational agents (ECAs), participants found the male agent more trustworthy than the female agent [48]. Conversely, in the context of soliciting donation with human-robot interaction, male participants attributed higher trust to an agent with a female voice than to a male agent voice [101]. These results could be explained by the stereotypical gender roles, where research has reported how females asserting dominance tend to be evaluated negatively (less friendly) compared to the male agent doing the same task [57], which explains the decreased trust in an airport screening context. However, in soliciting a donation, men are more likely than women to signal their status by conspicuously spending on luxurious and publicly observable goods, especially

in a romantic frame [196].

We argue that these findings show how the trust perception of CA gender is context-dependent. For example, in voice-user interaction, people found the product description more credible when the gender of the voice that read the description out loud matched with the stereotypical "gender" of the product [33 (p.26)]. Moreover, there was a finding that users were more likely to apply gender stereotypes when a chatbot operates within the gender-stereotypical subject domain, for example, taking a mechanic's role [101]. Thus, we expect to see a similar effect in our study that participants will perceive a specific gender of CA positively when the CA gender matches the context gender. In summary, our hypothesis for the RQ2 is:

- Hypothesis 2 (H2): Participants will show positive reliance, trust, intention to adopt, desire to cooperate, and perceived usability to CAs when CA gender correlates with context gender.

Gender-androgynous Agents

Previous studies with voice user interfaces (VUI) have shown that people unconsciously disapprove and dislike gender-androgynous voices, hypothetically, because it causes categorical tension [47, 33]. However, in text-based chatbot interaction, participants reported higher user satisfaction toward the non-gendered agent than either male or female chatbot [180]. As this study adopts a text-based chatbot agent, we also expect to see higher perceived usability with gender androgynous conditions. As there was a finding that perceived usability positively correlates to the self-reported intention to adopt score [197], we further hypothesize that gender androgynous agents will have a higher intention to adopt. In conclusion, our hypothesis for the RQ3 is:

- Hypothesis 3 (H3): Gender-androgynous agents will result in higher perceived usability, leading to higher intention to adopt.

Is Gender a Part of Metaphor?

As mentioned before, research has shown that perceived occupational competency was a better predictor of human trust than the robot gender in the context of human-robot interaction [192]. Moreover, robot gender least impacted the user trust compared to other attributes of robot acceptance, such as attitude towards the robot, affective evaluations, and cognitive evaluations [198]. Therefore, while we recognize the effect would exist based on the CA gender, we expect to see a bigger effect of metaphors with different perceived competence and warmth on our independent variables. To sum up, we propose our hypothesis for RQ4 as follows:

- Hypothesis 4 (H4): Metaphors will have a larger effect size in determining user trust, reliance, intention to adopt, desire to cooperate, and perceived usability than CA gender.

5.3 Method

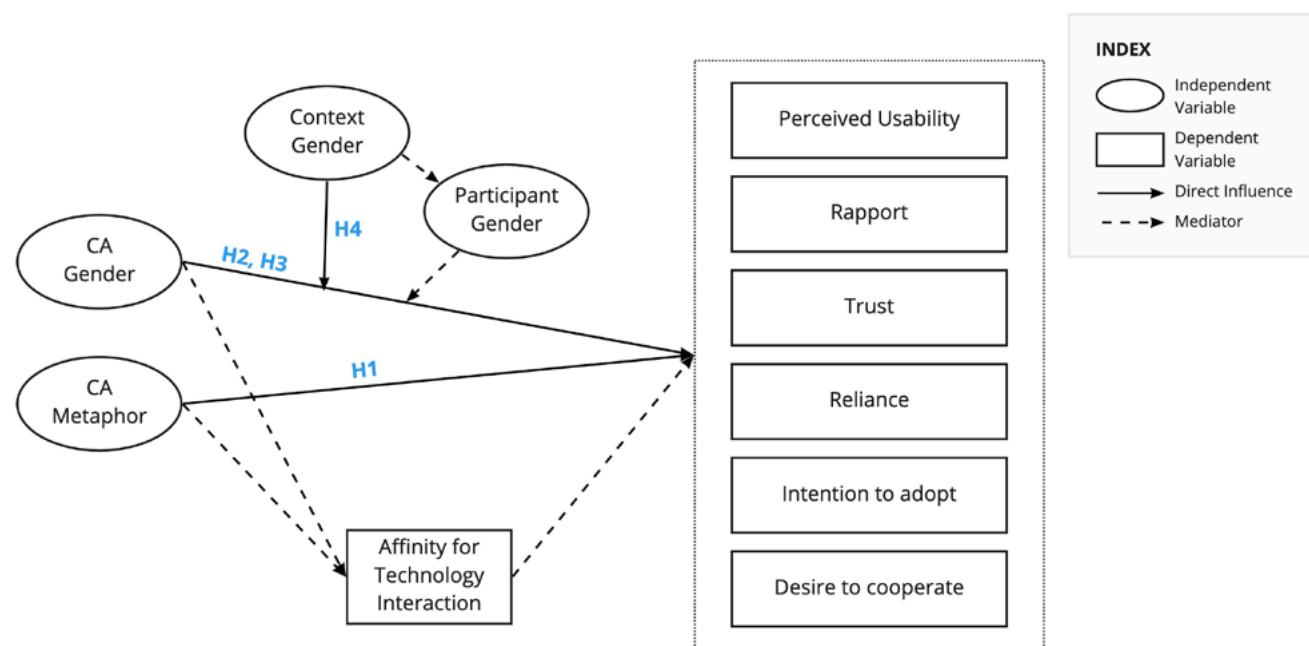


Figure 32. Research model diagram

We conducted between-subject factorial design (3 (CA metaphor) $\times$ 3 (CA gender) $\times$ 3 (“gendered” context)).

Study Design

Metaphor Design

We took metaphors with different perceived competence and warmth that previous research used: Trained Professional for high warmth and high competence (HW-HC) condition; Shrewd Sales Executive for low warmth and high competence (LW-HC); and Middle Schooler for high warmth and low competence (HW-LC) [69].

Gendered Agent Design

We adopted agent avatars that were validated and used in a previous study in Chapter 4. The avatars embody cartoon-like visual style in order to include our metaphor condition that varies in age. We also use the same name as we tested in Chapter 4 (female-Emma; male-Tom; androgynous-Jamie).

Using a gendered pronoun can help indicate gender. Several studies have reported that applying pronoun to a robot was successful in manipulating participants' perception towards the robot, even with gender-neutral physical look (NAO robot) [177]. Therefore, we refer three agent to participants using different gendered pronouns.

Product Gender

Based on previous research about product gender perceptions [199], we chose toothpaste for the gender-androgynous product, hairspray for the feminine product, and lawn mower for the masculine product. We then generated a fictional product list that contains at least 5 attributes. In order to prevent product image biasing participants, we decided to use black and white icon images for all the products (Figure 33).

Scenarios

Using the product list we composed, we designed two scenarios for each product gender category. Each

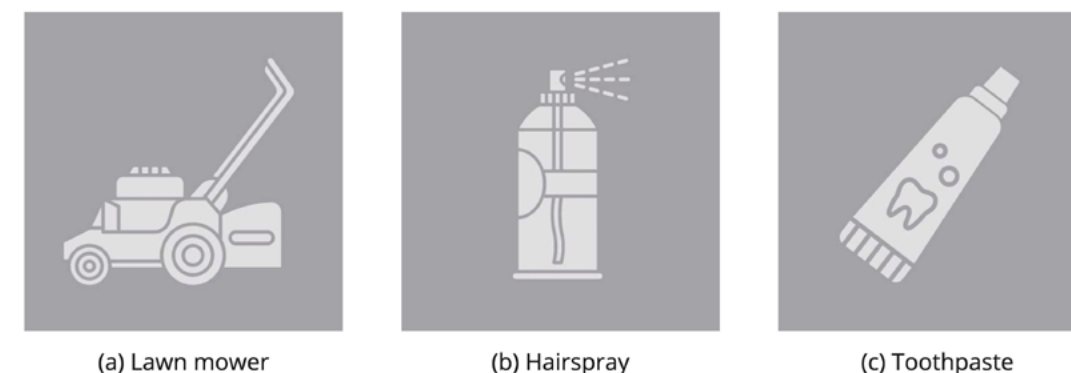


Figure 33. Product images

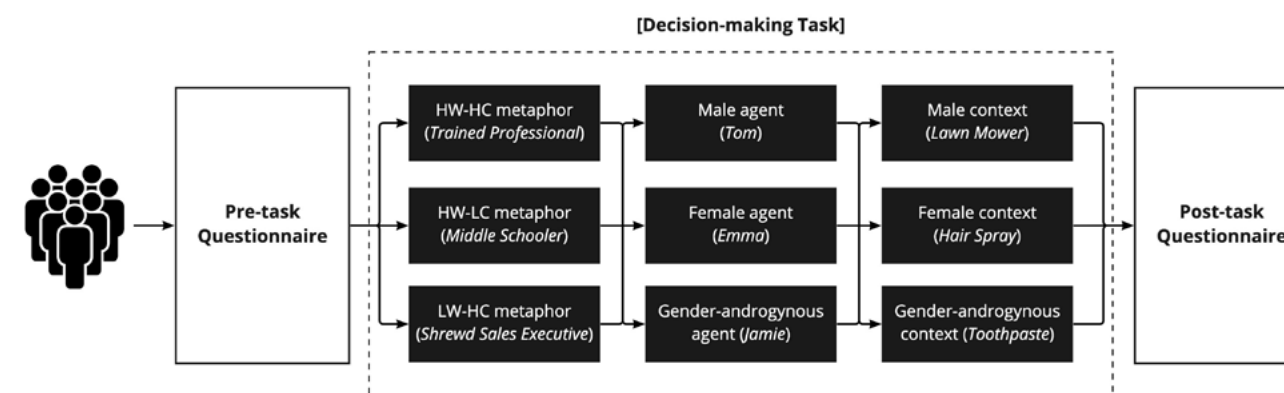


Figure 34. Study Procedure

product had varying levels of complexities, where the “Easy” scenario required participants to find a product that meets three constraints while the “Complex” scenario required five constraints. These scenario complexities were randomly shown to the participants to prevent question order bias.

Chatbot Recommendations

Similarly, CAs were designed to randomly recommend one correct and another incorrect recommendation. Therefore, participants could encounter the “Easy” scenario first and get incorrect recommendations, or get the “Complex” scenario first and get correct recommendations.

Measures

Gender demographic

We ask participants to indicate their gender by using the inclusive gender survey form suggested by Spiel et al. [58].

Affinity to Technology Interaction (ATI)

We used 9-item “Affinity for Technology Interaction” (ATI) questionnaire [200] based on a 6-point Likert scale from Completely Disagree to Completely Agree.

Perceived Usability

As many theoretical and empirical studies have suggested that perceived usability can support or prevent users’ interaction with the system [201], we measured perceived usability using surveys from previous studies [202, 203, 69]. Content of questions can be found in Appendix FIXME. Participants were asked to rate on a scale ranging from 1 (strongly disagree) to 7 (strongly agree).

CUI Rapport

As a conversational interface can trigger unique reactions for users to pick up social cues from the interface [29], we decide to measure the rapport between participants and CAs based on the validated “CRS-UX” model [204]. These question items were devised precisely for a conversational recommender system based on the established ResQue [205] met-

rics. We measured CUI Rapport on a 7-point Likert scale.

Trust

To measure perceived trust, we adopted domain-specific trust scales for online recommender agents [197]. From the metrics, we used trust specific subset of Trust-Competence, Trust-Benevolence, and Trust-Integrity. Questions can be found in Appendix A. We assessed these items on a 7-point scale, ranging from 1 (strongly disagree) to 7 (strongly agree).

Desire to Cooperate and Intention to Adopt

We borrow from previous work [69] to measure users’ self-reported desire to cooperate with the system, and their intention to adopt the chatbot. We measured this with a 7-point Likert scale, ranging from 1 (strongly likely) to 7 (strongly likely).

Procedure

Participants will start from the crowdsourcing platform Prolific. They will be redirected to a conversational interface that we built from HTML-based TickTalkTurk [206]. Figure 35 shows the flowchart of how this interface worked. After they finished the task with the given scenarios, they will be led to the Survey platform Qualtrics. Figure 34 shows an overview of the whole experiment process.

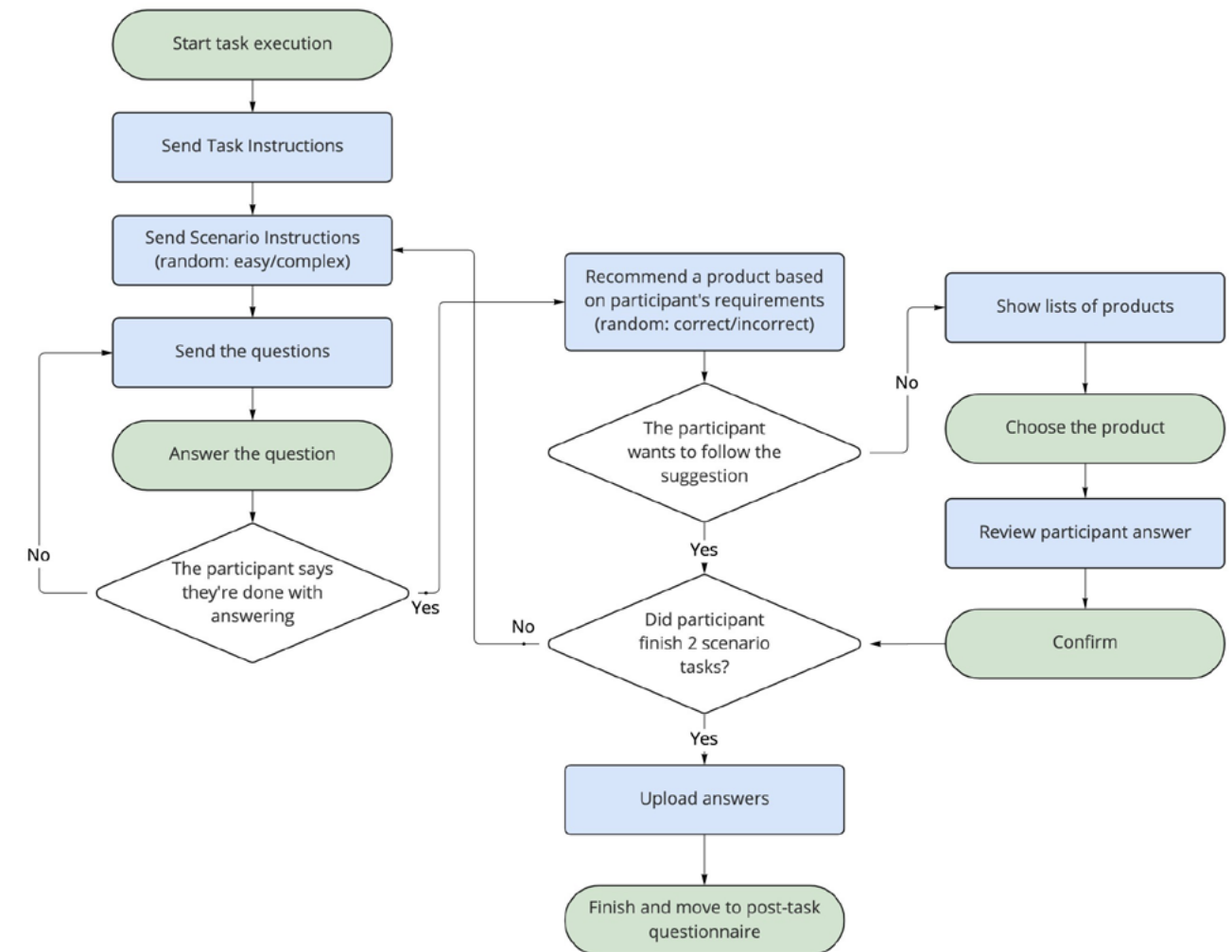


Figure 35. Chatbot interface flowchart

Thesis Discussion & Future Work

This chapter will give a summary of contributions on the graduation project. Afterwards, I give few design guidelines for designers in designing gender identity of CAs. Finally, some possible directions for further research will be identified.

- 6.1 Summary of Contributitons
- 6.2 Design Guidelines for CA Geder
- 6.3 Future Work

6.1. Summary of Contributions

We investigate ethical identity design of conversational agents in three angles: (1) developing framework to address ethical dilemma in CA identity design, (2) evaluating the effect of gender and metaphor in chatbot profile, and (3) calibrating gender stereotyping through metaphor manipulation. Each three perspectives is addressed by an individual chapter of this report.

Developing Framework to Address Ethical Dilemma in CA Identity Design

To address **RQ 3.1** and **RQ 3.2**, we developed an ethical framework by analyzing the ethical discussions in the conversational user interface (CUI) and HCI community. We presented the framework to an academic channel and synthesized valuable feedback.

While numerous voices have expressed the need for a better ethical framework in designing CA identity, few works have investigated this dilemma with an ethical framework. This work takes a needed first step in proposing the ethics framework for CA identity design. To this end, we reviewed and synthesized previous and recent literature on CA and gender stereotyping and identified the research gaps. Afterward, we define the layers of the social problem with CA identity design, which designers can use to articulate the reason behind the dilemma. Finally, we introduce dialogical ethics that can be used as a critical thinking guide for designers facing the CA identity design dilemmas.

Evaluating the Effect of Gender and Metaphor in Chatbot Profile

To address **RQ 4.1** and **RQ 4.2**, we conducted a mixed-method study to get quantitative and qualitative insight into chatbot profiles' gender cues and metaphors.

We show that the effect of textual metaphor proven in previous literature persisted in its impact on perception, even after the visual cue (avatar) and gender cues are attached. We then collected evidence on how metaphor had a larger effect on a statistically significant difference in perceived warmth and competence. Finally, we provided qualitative themes that impact users' first impressions of the CA profile.

Calibrate Gender Stereotyping through Metaphor Manipulation

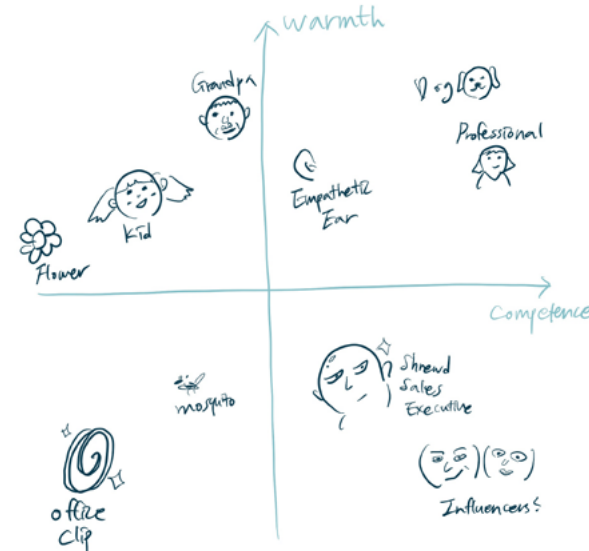
To address **RQ 5.1**, **RQ 5.2**, **RQ 5.3**, and **RQ 5.4**, we conducted an empirical study on the impact of the CA metaphor with different gender cues (female, male, androgynous). We specifically investigated this in the context of human decision-making with the CA recommender system.

**Important to note: This part of the thesis is a work in progress. We are planning to publish the result in the academic venue. The thesis will be updated accordingly within a year after the thesis publication date on TU Delft Repository.*

6.2 Design Guidelines

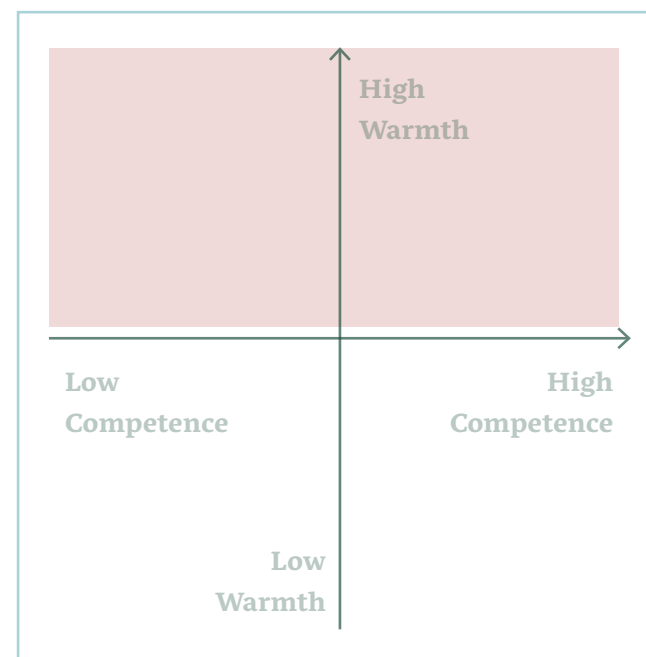
Brainstorm your metaphor in the Warmth-Competence axis

Prepare a blank paper and draw warmth and competence axis. Now, brainstorm any metaphors that comes up in your mind that suits the context of the CA you are designing for.



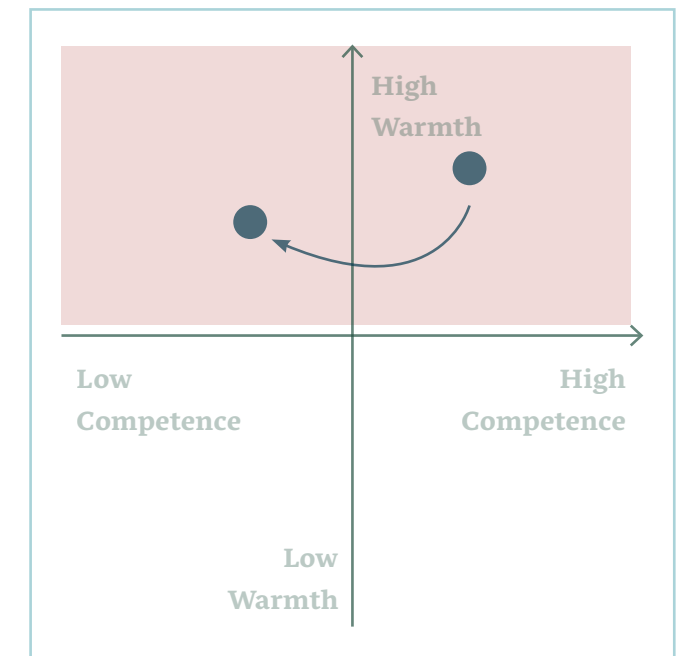
Try to Stick to high-warmth metaphor

Try to stick to high-warmth metaphor in order to induce more positive reaction from the users, especially to improve their intention to adopt and desire to cooperate [69].



Be aware of the trade-offs of using high-competence metaphor

Using high-competence metaphor will intrigue people to try out the CA as it sparks curiosity. However, if people find out the limitations in conversational ability of CAs, they will quickly be disappointed and their intention to adopt and desire to cooperate will drop. Therefore, try to calibrate the competence markers well by communicating the system ability once user starts their journey.



Using an adjective can help in forming the metaphor

We found that people pick up the impression of adjective quite sensitively. Validate the impression of an adjective you want to use in the CA metaphor with potential users to see if it induces similar reaction that you want from users.

Outstanding Inspiring Smart
Beautiful helpful Confident Amazing devoted
lazy Friendly Considerate mature
Sensible Shy Cute eager
timid Pretty Responsible
Handsome naughty Likeable
clever average brilliant

Introduction messages matter

Pay attention to how CAs greet users, as it can impact first impressions more than you might think. If you are aiming for high-warmth, greet them with warming welcome.



Pick metaphors that matches the task characteristic

Do not just focus on warmth and competence level, but also pay attention if the metaphor you picked can be relevant to the tasks you are assigning CA to. People will bring their real-world experiences when they interact with your CA.

6.3 Future Work

This thesis investigates how designers can make more cautious design choices when it comes to the gender identity design of the CA. In this regard, we presented an ethics framework for designers to use when navigating the dilemma. In addition, we proposed to calibrate the effect of CA gender by manipulating the metaphor. While we believe our results are promising, there are few directions that future works can investigate.

Developing Design Tool using the Dialogical Ethics

While we analyzed the layers that cause the dilemma in CA gender design and proposed dialogical ethics to navigate through this space, it is still in a theoretical phase where it's hard to be immediately deployed in design practice. Therefore, future work can develop a more practical tool for designers to adopt dialogical ethics.

How to design gender-androgynous cues

In this thesis, we reported the impact of gender-androgynous agents in the user evaluations and compared it to other gendered agents. However, it is still a vague area to design an inclusive visual avatar that can be truly perceived as androgynous. As people develop their gender stereotyping from a very early age, previous research has found how people assign the agent to either female or male, even when designers intended the agent to be perceived as androgynous. Therefore, it will be exciting research to study the effect of gender-neutral markers in CA representations.

Does metaphor persist in Voice User Interface?

We only focused on text-based task-oriented chatbots. However, as people sensitively react to the

gender of CA's voice varying in pitch and tone, it will be an interesting research opportunity to see if metaphor will still "overpower" the effect of gender.

Varying metaphors in different cultural contexts

Metaphors are perceived very differently based on the cultural background of the people understanding them. A good example is a voice gender design of a Japanese telephone-based stock brokerage system, where users are introduced to stock information in a female voice while they complete a purchase with male voice agents. This is derived from the Japanese custom in which females are considered more suited to researching and communicating the stock information, while males are suited to handle the purchase. This is in contrast to the case of the US, where males stereotypically are considered more suitable for the job.

However, in this thesis, we only focused on US participants. Therefore, future research can investigate and compare the difference in the perceptions of the metaphors we used in this work.

References

- [1] Eiraku Maiko. 2019. A funeral for dead robot dogs. *NHK World-Japan*. https://www3.nhk.or.jp/nhkworld/en/news/backstories/346/#::~:tex-t=On%20the%20altar%20at%20Kofukuji,a%20funeral%20for%20a%20human_.
- [2] Reeves, B. and Nass, C., 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10, p.236605.
- [3] Nass, C., Steuer, J. and Tauber, E.R., 1994, April. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 72-78).
- [4] Alan M Turing. 1950. Computing machinery and intelligence. *Mind* 59, 236 (1950), 433–460.
- [5] Karray, F., Alemzadeh, M., Abou Saleh, J. and Arab, M.N., 2017. Human-computer interaction: Overview on state of the art. *International journal on smart sensing and intelligent systems*, 1(1).
- [6] Owen, C., 2007. Design thinking: Notes on its nature and use. *Design Research Quarterly*, 2(1), pp.16-27.
- [7] Ashby, S., Hanna, J., Matos, S., Nash, C. and Faria, A., 2019, November. Fourth-wave HCI meets the 21st century manifesto. In *Proceedings of the Halfway to the Future Symposium 2019* (pp. 1-11).
- [8] Blevis, E., Chow, K., Koskinen, I., Poggenpohl, S. and Tsin, C., 2014. Billions of interaction designers. *Interactions*, 21(6), pp.34-41.
- [9] Johnson, R.R., 1998. User-centered technology: A rhetorical theory for computers and other mundane artifacts. SUNY press.
- [10] Putnam, C., Reiner, A., Ryou, E., Caputo, M., Cheng, J., Allen, M. and Singamaneni, R., 2016. Human-centered design in practice: Roles, definitions, and communication. *Journal of Technical Writing and Communication*, 46(4), pp.446-470.
- [11] Himma, K.E. and Tavani, H.T. eds., 2008. The handbook of information and computer ethics (pp. 135-156). Hoboken: Wiley.
- [12] Friedman, B., Hendry, D.G. and Borning, A., 2017. A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction*, 11(2), pp.63-125.
- [13] Fessler, L., 2017. We tested bots like Siri and Alexa to see who would stand up to sexual harassment. [online] Quartz. Available at: <<https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/>> [Accessed 25 June 2022].
- [14] Deahl, Dani. 2017. "Samsung Adds And Swiftly Removes Sexist Bixby Descriptor Tags". The Verge. <https://www.theverge.com/2017/7/19/15998668/samsung-adds-removes-sexist-bixby-descriptor-tags>.
- [15] Dahlbäck, N., Jönsson, A. and Ahrenberg, L., 1993. Wizard of Oz studies—why and how. *Knowledge-based systems*, 6(4), pp.258-266.
- [16] Schön, D.A., 1979. Generative metaphor: A perspective on problem-setting in social policy. *Metaphor and thought*, 2, pp.137-163.
- [17] Saffer, D., 2005. The role of metaphor in interaction design. *Information Architecture Summit*, 6.
- [18] Cila, N., 2013. Metaphors we design by: The use of metaphors in product design.
- [19] Purington, A., Taft, J.G., Sannon, S., Bazarova, N.N. and Taylor, S.H., 2017, May. “ Alexa is my new BFF” Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems* (pp. 2853-2859).
- [20] Gao, Y., Pan, Z., Wang, H. and Chen, G., 2018, October. Alexa, my love: analyzing reviews of amazon echo. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (pp. 372-380). IEEE.
- [21] Karjalainen, K.D., Romell, A.E.S., Ratsamee, P., Yantac, A.E., Fjeld, M. and Obaid, M., 2017, October. Social drone companion for the home environment: A user-centric exploration. In *Proceedings of the 5th International Conference on Human Agent Interaction* (pp. 89-96).
- [22] Jannach, D., Manzoor, A., Cai, W. and Chen, L., 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5), pp.1-36.
- [23] Christakopoulou, K., Radlinski, F. and Hofmann, K., 2016, August. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 815-824).
- [24] Sun, Y. and Zhang, Y., 2018, June. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 235-244).
- [25] Bergen, H., 2016. ‘I’d blush if I could’: Digital assistants, disembodied cyborgs and the problem of gender. *Word and Text, A Journal of*

- Literary Studies and Linguistics*, 6(01), pp.95-113.
- [26] Ni Loideain, N. and Adams, R., 2018. From Alexa to Siri and the GDPR: the gendering of virtual personal assistants and the role of EU data protection law. *King’s College London Dickson Poon School of Law Legal Studies Research Paper Series*.
- [27] West, M., Kraut, R. and Chew, H.E., 2019. I’d blush if I could: Closing gender divides in digital skills through education. *UNESCO for the EQUALS Skills Coalition*.
- [28] Feine, J., Gnewuch, U., Morana, S. and Maedche, A., 2019, November. Gender bias in chatbot design. In *International Workshop on Chatbot Research and Design* (pp. 79-93). Springer, Cham.
- [29] Feine, J., Gnewuch, U., Morana, S. and Maedche, A., 2019. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, 132, pp.138-161.
- [30] Lee, E.J., Nass, C. and Brave, S., 2000, April. Can computer-generated speech have gender? An experimental test of gender stereotype. In *CHI’00 extended abstracts on Human factors in computing systems* (pp. 289-290).
- [31] Pak, R., McLaughlin, A.C. and Bass, B., 2014. A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults. *Ergonomics*, 57(9), pp.1277-1289.
- [32] Strand, E.A., 2000. Gender stereotype effects on speech processing. *The Ohio State University*.
- [33] Nass, C.I. and Brave, S., 2005. *Wired for speech: How voice activates and advances the human-computer relationship* (p. 9). Cambridge: MIT press.
- [34] Nass, C. and Moon, Y., 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1), pp.81-103.
- [35] Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A.Y. and Coiera, E., 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9), pp.1248-1258.
- [36] Veletsianos, G. and Russell, G.S., 2014. Pedagogical agents. In *Handbook of research on educational communications and technology* (pp. 759-769). Springer, New York, NY.
- [37] Winkler, R., Hobert, S., Salovaara, A., Söllner, M. and Leimeister, J.M., 2020, April. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).
- [38] Guo, J., Guo, J., Yang, C., Wu, Y. and Sun, L., 2021, May. Shing: A Conversational Agent to Alert Customers of Suspected Online-payment Fraud with Empathetical Communication Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-11).
- [39] McTear, M.F., Callejas, Z. and Griol, D., 2016. The conversational interface (Vol. 6, No. 94, p. 102). Cham: Springer.
- [40] Zheng, Q., Tang, Y., Liu, Y., Liu, W. and Huang, Y., 2022, April. UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library. In *CHI Conference on Human Factors in Computing Systems* (pp. 1-24).
- [41] Bittner, E.A., Oeste-Reiß, S. and Leimeister, J.M., 2019. Where is the bot in our team? Toward a taxonomy of design option combinations for conversational agents in collaborative work. In *Hawaii International Conference on System Sciences (HICSS)*.
- [42] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*. Springer, 946–956.
- [43] Morten Johan Mygland, Morten Schibbye, Ilias O Pappas, and Polyxeni Vassi-Iakopoulou. 2021. Afordances in human-chatbot interaction: a review of the literature. In *Conference on e-Business, e-Services and e-Society*. Springer, 3–17.
- [44] Andreea Danielescu. 2020. Eschewing gender stereotypes in voice assistants to promote inclusion. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–3.
- [45] Charles Hannon. 2016. Gender and status in voice user interfaces. *Interactions* 23, 3 (2016), 34–37.
- [46] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology* 27, 10 (1997), 864–876.
- [47] Lopatovska, I., Brown, D. and Korshakova, E., 2022, February. Contextual Perceptions of Feminine-, Masculine-and Gender-Ambiguous-Sounding Conversational Agents. In *International Conference on Information* (pp. 459-480). Springer, Cham.

[48] Jay F Nunamaker, Douglas C Derrick, Aaron C Elkins, Judee K Burgoon, and Mark W Patton. 2011. Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems* 28, 1 (2011), 17–48.

[49] Kervyn, N., Fiske, S.T. and Malone, C., 2012. Brands as intentional agents framework: How perceived intentions and ability can map brand perception. *Journal of Consumer Psychology*, 22(2), pp.166-176.

[50] Aaker, J., Vohs, K.D. and Mogilner, C., 2010. Nonprofits are seen as warm and for-profits as competent: Firm stereotypes matter. *Journal of Consumer Research*, 37(2), pp.224-237.

[51] Søndergaard, M.L.J. and Hansen, L.K., 2018, June. Intimate futures: Staying with the trouble of digital personal assistants through design fiction. In *Proceedings of the 2018 designing interactive systems conference* (pp. 869-880).

[52] Luciano Floridi. 2010. *The Cambridge handbook of information and computer ethics*. Cambridge University Press.

[53] Sheryl Brahnam and Antonella De Angeli. 2012. Gender affordances of conversational agents. *Interacting with Computers* 24, 3 (2012), 139–153.

[54] Silvervarg, A., Raukola, K., Haake, M. and Gulz, A., 2012, September. The effect of visual gender on abuse in conversation with ECAs. In *International conference on intelligent virtual agents* (pp. 153-160). Springer, Berlin, Heidelberg.

[55] Magnus Haake and Agneta Gulz. 2008. Visual stereotypes and virtual pedagogical agents. *Journal of Educational Technology & Society* 11, 4 (2008), 1–15.

[56] George Veletsianos, Cassandra Scharber, and Aaron Doering. 2008. When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with computers* 20, 3 (2008), 292–301.

[57] Friederike Eyssel and Frank Hegel. 2012. (s) he’s got the look: Gender stereotyping of robots 1. *Journal of Applied Social Psychology* 42, 9 (2012), 2213–2230.

[58] Spiel, K., Haimson, O.L. and Lottridge, D., 2019. How to do better with gender on surveys: a guide for HCI researchers. *Interactions*, 26(4), pp.62-65.

[59] Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. 2021. “It’s Complicated”: Negotiating Accessibility and (Mis) Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[60] Simone Stumpf, Anicia Peters, Shaowen Bardzell, Margaret Burnett, Daniela Busse, Jessica Cauchard, and Elizabeth Churchill. 2020. Gender-inclusive HCI research and design: A conceptual review. *Foundations and Trends in Human–Computer Interaction* 13, 1 (2020), 1–69.

[61] Marilyn B Brewer and Layton N Lui. 1989. The primacy of age and sex in the structure of person categories. *Social cognition* 7, 3 (1989), 262–274.

[62] Susan T Fiske and Steven L Neuberg. 1990. A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology*. Vol. 23. Elsevier, 1–74.

[63] Alice H Eagly, Anne E Beall, and Robert J Sternberg. 2005. *The psychology of gender*. Guilford Press.

[64] Judith A Howard and Jocelyn A Hollander. 1997. Gendered situations, gendered selves: A gender lens on social psychology. Vol. 2. Rowman & Littlefield.

[65] JA Bargh. 1999. The cognitive monster: The case against controllability of automatic stereotype effects.(In) Chaiken, S., & Trope, Y.(Eds.), *Dual Process Theories in Social Psychology*. Guilford Press 361 (1999), 363.

[66] Patricia G Devine. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology* 56, 1 (1989), 5.

[67] Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology* 40 (2008), 61–149.

[68] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowd-sourcing. In *CHI Conference on Human Factors in Computing Systems (CHI ’22)*, April 29-May 5, 2022, New Orleans, LA, USA. ACM New York, NY, USA, 32 pages.

[69] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of

human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26

[70] Sylvie Borau, Tobias Otterbring, Sandra Laporte, and Samuel Fosso Wamba. 2021. The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology and Marketing* 38, 7 (2021), 1052–1068. <https://doi.org/10.1002/mar.21480>

[71] Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J.S., Leimeister, J.M.M. and Bernstein, A., 2021, May. Female by Default?—Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-7).

[72] Emmie Matsuno and Stephanie L Budge. 2017. Non-binary/genderqueer identities: A critical review of the literature. *Current Sexual Health Reports* 9, 3 (2017), 116–120.

[73] Donald A Norman. 1986. Cognitive engineering. *User centered system design* 31 (1986), 61.

[74] Alexander Kunst. 2019. Preferences for male or female voices for digital voice assistant among residents of the United States, as of April 2017. (2019). <https://www.statista.com/statistics/702911/united-states-digital-voice-assistants-survey-gender-preferences/>

[75] Julia Driver. 2014. The History of Utilitarianism. In *The Stanford Encyclopedia of Philosophy* (Winter 2014 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[76] ter Stal, S., Tabak, M., op den Akker, H., Beinema, T. and Hermens, H., 2020. Who do you prefer? The effect of age, gender and role on users’ first impressions of embodied conversational agents in eHealth. *International Journal of Human–Computer Interaction*, 36(9), pp.881-892.

[77] Michael Ahmadi, Rebecca Eilert, Anne Weibert, Volker Wulf, and Nicola Marsden. 2020. “ We want to push the industry via communication”... Designing Communication Measures to Foster Gender Diversity in a Video Game Company. *Proceedings of the ACM on Human-Computer Interaction* 4, GROUP (2020), 1–26.

[78] Xeniya Kondrat et al. 2015. Gender and video games: How is female gender generally represented in various genres of video games? *Journal of comparative research in anthropology and sociology* 6, 01 (2015), 171–193.

[79] Pragmatic Health Ethics Research Unit. 2020. Reflections on ethical deliberation and the dialogical approach at the heart of healthcare. <https://youtu.be/- uNpYmw7jG8>

[80] Alison Adam. 2005. *Gender, ethics and information technology*. Springer.

[81] Daryl Koehn. 2012. *Rethinking feminist ethics: Care, trust and empathy*. Routledge.

[82] Albert C Molewijk, Tineke Abma, Margreet Stolper, and Guy Widdershoven. 2008. Teaching ethics in the clinic. The theory and practice of moral case deliberation. *Journal of Medical Ethics* 34, 2 (2008), 120–124.

[83] Margreet Stolper, Bert Molewijk, and Guy Widdershoven. 2016. Bioethics education in clinical settings: theory and practice of the dilemma method of moral case deliberation. *BMC Medical Ethics* 17, 1 (2016), 1–10.

[84] Albert C Molewijk, Tineke Abma, Margreet Stolper, and Guy Widdershoven. 2008. Teaching ethics in the clinic. The theory and practice of moral case deliberation. *Journal of Medical Ethics* 34, 2 (2008), 120–124.

[85] Froukje C Weidema, Bert AC Molewijk, Frans Kamsteeg, and Guy AM Widdershoven. 2013. Aims and harvest of moral case deliberation. *Nursing Ethics* 20, 6 (2013), 617–631.

[86] Lieke Josephina Jeanne Johanna Vrouwenraets, Laura A Hartman, Irma M Hein, Annelou LC de Vries, Martine C de Vries, and Bert AC Molewijk. 2020. Dealing with moral challenges in treatment of transgender children and adolescents: evaluating the role of Moral Case Deliberation. *Archives of Sexual Behavior* 49, 7 (2020), 2619–2634.

[87] Stephanie Carnell and Benjamin Lok. 2018. The effects of author identity on dialogue for virtual human communication skills training. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 65–70.

[88] Caroline Criado Perez. 2019. *Invisible women: Data bias in a world designed for men*. Abrams.

[89] Catherine D’Ignazio, Alexis Hope, Becky Michelson, Robyn Churchill, and Ethan Zuckerman. 2016. A Feminist HCI Approach to Designing Postpartum Technologies: “ When I first saw a breast pump I was wondering if it was a joke”. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2612–2622.

[90] Amsterdam ELS ROMMES, Ellen Van Oost, and Nelly Oudshoorn. 2005. Gender in the Design of the Digital City of Amsterdam. In *Virtual Gender*. Routledge, 263–284.

- [91] Rien M JPA Janssens, Ezra Van Zadelhoff, Ger Van Loo, Guy AM Widdershoven, and Bert AC Molewijk. 2015. Evaluation and perceived results of moral case deliberation: A mixed methods study. *Nursing Ethics* 22, 8 (2015), 870–880.
- [92] Lieke Josephina Jeanne Johanna Vrouenraets, Laura A Hartman, Irma M Hein, Annelou LC de Vries, Martine C de Vries, and Bert AC Molewijk. 2020. Dealing with moral challenges in treatment of transgender children and adolescents: evaluating the role of Moral Case Deliberation. *Archives of Sexual Behavior* 49, 7 (2020), 2619–2634.
- [93] Leikas, J., Koivisto, R. and Gotcheva, N., 2019. Ethical framework for designing autonomous intelligent systems. *Journal of Open Innovation: Technology, Market, and Complexity*, 5(1), p.18.
- [94] Wagner, K., Nimmermann, F. and Schramm-Klein, H., 2019, January. Is it human? The role of anthropomorphism as a driver for the successful acceptance of digital voice assistants. In *proceedings of the 52nd Hawaii international conference on system sciences*.
- [95] Waytz, A., Heafner, J. and Epley, N., 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology*, 52, pp.113-117.
- [96] McMillan, D. and Jaber, R., 2021, July. Leaving the Butler Behind: The Future of Role Reproduction in CUI. In *CUI 2021-3rd Conference on Conversational User Interfaces* (pp. 1-4).
- [97] Anna-Maria Meck, Justin Edwards, Diego Garaialde, Marion Bartl, Philip Doyle, and Leigh Clark. 2022. Design for the User You Want, Not the User You Have?. *CUI@CHI2022 Workshop position paper*. https://www.conversationaluserinterfaces.org/workshops/CHI2022/pdfs/meck_Bias_Workshop_Paper.pdf
- [98] Porcheron, Martin, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. "Voice interfaces in everyday life." In *proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1-12. 2018.
- [99] Luger, Ewa, and Abigail Sellen. "" Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents." In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 5286-5297. 2016.
- [100] Large, David R., Leigh Clark, Gary Burnett, Kyle Harrington, Jacob Luton, Peter Thomas, and Pete Bennett. "" It's small talk, jim, but not as we know it." engendering trust through human-agent conversation in an autonomous, self-driving car." In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pp. 1-7. 2019.
- [101] Siegel, Mikey, Cynthia Breazeal, and Michael I. Norton. "Persuasive robotics: The influence of robot gender on human behavior." In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2563-2568. IEEE, 2009.
- [102] Rheu, Minjin, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. "Systematic review: Trust-building factors and implications for conversational agent design." *International Journal of Human–Computer Interaction* 37, no. 1 (2021): 81-96.
- [103] Tolmeijer, Suzanne, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. "Second chance for a first impression? Trust development in intelligent system interaction." In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*, pp. 77-87. 2021.
- [104] Eagly, Alice H. *Sex differences in social behavior: A social-role interpretation*. Psychology Press, 2013.
- [105] Eagly, Alice H., and Antonio Mladinic. "Gender stereotypes and attitudes toward women and men." *Personality and social psychology bulletin* 15, no. 4 (1989): 543-558.
- [106] Glick, Peter, and Susan T. Fiske. "Sexism and other" isms": Independence, status, and the ambivalent content of stereotypes." (1999).
- [107] Tannenbaum, Cara, Robert P. Ellis, Friederike Eyssel, James Zou, and Londa Schiebinger. "Sex and gender analysis improves science and engineering." *Nature* 575, no. 7781 (2019): 137-146.'
- [108] Wang, Yan, and James E. Young. "Beyond “pink” and “blue”: gendered attitudes towards robots in society." *GenderIT* 2014 (2014).
- [109] Epley, Nicholas, Adam Waytz, and John T. Cacioppo. "On seeing human: a three-factor theory of anthropomorphism." *Psychological review* 114, no. 4 (2007): 864.
- [110] McDonnell, Marian, and David Baxter. "Chatbots and gender stereotyping." *Interacting with Computers* 31, no. 2 (2019): 116-121.
- [111] McTear, Michael F. "The rise of the conversational interface: A new kid on the block?." In *International workshop on future and emerging trends in language technology*, pp. 38-49. Springer, Cham, 2016.
- [112] Weizenbaum, Joseph. "ELIZA—a computer program for the study of natural language communication between man and machine." *Com-*

- munications of the ACM* 9, no. 1 (1966): 36-45.
- [113] Mikic, Fernando A., Juan C. Burguillo, Martín Llamas, Daniel A. Rodríguez, and Eduardo Rodríguez. "Charlie: An aiml-based chatterbot which works as an interface among ines and humans." In *2009 EAEEIE Annual Conference*, pp. 1-6. IEEE, 2009.
- [114] Klopfenstein, Lorenz Cuno, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. "The rise of bots: A survey of conversational interfaces, patterns, and paradigms." In *Proceedings of the 2017 conference on designing interactive systems*, pp. 555-565. 2017.
- [115] Gambino, Andrew, Jesse Fox, and Rabindra A. Ratan. "Building a stronger CASA: Extending the computers are social actors paradigm." *Human-Machine Communication* 1 (2020): 71-85.
- [116] Rapp, Amon, Lorenzo Curti, and Arianna Boldi. "The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots." *International Journal of Human-Computer Studies* 151 (2021): 10263
- [117] Kim Y, Sundar SS. Anthropomorphism of computers: is it mindful or mindless? *Computers in Human Behavior* 2012; 28:241–250.
- [118] Kim, Ahyeon, Minha Cho, Jungyong Ahn, and Yongjun Sung. "Effects of gender and relationship type on the response to artificial intelligence." *Cyberpsychology, Behavior, and Social Networking* 22, no. 4 (2019): 249-253.
- [119] Li, Mengjun, and Ayoung Suh. "Machinelike or humanlike? A literature review of anthropomorphism in AI-enabled technology." (2021).
- [120] Clark, Leigh, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane et al. "What makes a good conversation? Challenges in designing truly conversational agents." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-12. 2019.
- [121] Brabra, Hayet, Marcos Báez, Boualem Benatallah, Walid Gaaloul, Sara Bouguelia, and Shayan Zamanirad. "Dialogue management in conversational systems: a review of approaches, challenges, and opportunities." *IEEE Transactions on Cognitive and Developmental Systems* (2021).
- [122] Zhang, Zheng, Minlie Huang, Zhongzhou Zhao, Feng Ji, Haiqing Chen, and Xiaoyan Zhu. "Memory-augmented dialogue management for task-oriented dialogue systems." *ACM Transactions on Information Systems (TOIS)* 37, no. 3 (2019): 1-30.
- [123] Strohkorb Sebo, Sarah, Margaret Traeger, Malte Jung, and Brian Scassellati. "The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams." In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 178-186. 2018.
- [124] Wang, Ning, David V. Pynadath, and Susan G. Hill. "Trust calibration within a human-robot team: Comparing automatically generated explanations." In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 109-116. IEEE, 2016.
- [125] Martelaro, Nikolas, Victoria C. Nneji, Wendy Ju, and Pamela Hinds. "Tell me more designing HRI to encourage more trust, disclosure, and companionship." In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 181-188. IEEE, 2016.
- [126] Torre, Ilaria, Jeremy Goslin, Laurence White, and Debora Zanatto. "Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience." In *Proceedings of the Technology, Mind, and Society*, pp. 1-6. 2018.
- [127] Gong, Li, and Clifford Nass. "When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference." *Human communication research* 33, no. 2 (2007): 163-193.
- [128] Qiu, Lingyun, and Izak Benbasat. "Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems." *Journal of management information systems* 25, no. 4 (2009): 145-182.
- [129] Qiu, Sihang, Ujwal Gadiraju, and Alessandro Bozzon. "Estimating conversational styles in conversational microtask crowdsourcing." *Proceedings of the ACM on Human-Computer Interaction* 4, no. CSCW1 (2020): 1-23.
- [130] Colburn, Timothy R., and Gary M. Shute. "Metaphor in computer science." *Journal of applied logic* 6, no. 4 (2008): 526-533.
- [131] Richard Mander, Gitta Salomon, and Yin Yin Wong. 1992. A “pile” metaphor for supporting casual organization of information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 627–634.
- [132] Robin Sease. 2008. Metaphor’s role in the information behavior of humans interacting with computers. *Information technology and libraries* 27, 4 (2008), 9–16.
- [133] Colburn, Timothy R., and Gary M. Shute. "Metaphor in computer science." *Journal of applied logic* 6, no. 4 (2008): 526-533.
- [134] Susan A Gelman and Cristine H Legare. 2011. Concepts and folk theories. *Annual review of anthropology* 40 (2011), 379–398.

- [135] Jonas Oppenlaender. 2020. Socially Augmented Crowdsourced Collection of Folk Theories. (2020).
- [136] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In Proceedings of the 2016 CHI conference on human factors in computing systems. 2371–2382.
- [137] Megan French and Jeff Hancock. 2017. What's the Folk Theory? Reasoning About Cyber-Social Systems. SSRN Electronic Journal (2017), 1–37. <https://doi.org/10.2139/ssrn.2910571>
- [138] IgnacioSiles,AndrésSegura-Castillo,RicardoSolís,andMónicaSancho.2020. Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global South. *Big Data & Society* 7, 1 (2020), 2053951720923377.
- [139] Sari Kujala, Ruth Mugge, and Talya Miron-Shatz. 2017. The role of expectations in service evaluation: A longitudinal study of a proximity mobile payment service. *International Journal of Human-Computer Studies* 98 (2017), 51–61.
- [140] Lakoff, George, and Mark Johnson. Metaphors we live by. University of Chicago press, 2008.
- [141] Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82, 878–902.
- [142] Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619– 619. doi:10.1126/science.1134475
- [143] Hwang, Angel Hsing-Chi, and Andrea Stevenson Won. "AI in Your Mind: Counterbalancing Perceived Agency and Experience in Human-AI Interaction." In CHI Conference on Human Factors in Computing Systems Extended Abstracts, pp. 1-10. 2022.
- [144] Bakan, D. (1966). The duality of human existence: Isolation and communion in Western man. Boston: Beacon Press
- [145] Abele, Andrea E., Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. "Facets of the fundamental content dimensions: Agency with competence and assertiveness—Communion with warmth and morality." *Frontiers in psychology* 7 (2016): 1810.
- [146] Zohar Gilad, Ofra Amir, and Liat Levontin. 2021. The Effects of Warmth and Competence Perceptions on Users' Choice of an AI System. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 1–13.
- [147] Minha Lee, Dimosthenis Kontogiorgos, Ilaria Torre, Michal Luria, Ravi Tejwani, Matthew J. Dennis, and Andre Pereira. 2021. Robo-Identity: Exploring Artificial Identity and Multi-Embodiment. In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (Boulder, CO, USA) (HRI '21 Companion). Association for Computing Machinery, New York, NY, USA, 718–720. <https://doi.org/10.1145/3434074.3444878>
- [148] Minha Lee, Peter Ruijten, Lily Frank, Yvonne de Kort, and Wijnand IJsselsteijn. 2021. People May Punish, But Not Blame Robots. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 715, 11 pages. <https://doi.org/10.1145/3411764.3445284>
- [149] Cikara, Mina, Rachel A. Farnsworth, Lasana T. Harris, and Susan T. Fiske. "On the wrong side of the trolley track: Neural correlates of relative social valuation." *Social cognitive and affective neuroscience* 5, no. 4 (2010): 404-413.
- [150] Anand, Priya. 2021. "Bloomberg - Are You A Robot?". Bloomberg.Com. <https://www.bloomberg.com/news/articles/2021-12-22/amazon-s-voice-controlled-smart-speaker-alexa-can-t-hold-customer-interest-docs>.
- [151] Kuczewski, Mark. "Casuistry and principlism: the convergence of method in biomedical ethics." *Theoretical Medicine and Bioethics* 19, no. 6 (1998): 509-524.
- [152] Ruane, Elayne, Abeba Birhane, and Anthony Ventresque. "Conversational AI: Social and Ethical Considerations." In AICS, pp. 104-115. 2019.
- [153] Minha Lee, Gale Lucas, and Jonathan Gratch. 2021. Comparing mind perception in strategic exchanges: Human-agent negotiation, dictator and ultimatum games. *Journal of Multimodal User Interfaces*. In press (2021), 1–15.
- [154] Lee, Minha, Gale Lucas, Johnathan Mell, Emmanuel Johnson, and Jonathan Gratch. "What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations." In Proceedings of the 19th ACM international conference on intelligent virtual agents, pp. 38-45. 2019.
- [155] Hildebrand, C., Bergner, A.: Conversational robo advisors as surrogates of trust: onboarding experience, firm perception, and consumer financial decision making. *J. Acad. Mark. Sci.* (2020). <https://doi.org/10.1007/s11747-020-00753-z>.
- [156] Bond, Raymond R., Maurice D. Mulvenna, Hui Wan, Dewar D. Finlay, Alexander Wong, Ansgar Koene, Rob Brisk, Jennifer Boger, and

- Tameem Adel. "Human Centered Artificial Intelligence: Weaving UX into Algorithmic Decision Making." In RoCHI, pp. 2-9. 2019.
- [157] Awad, E. et al. The Moral Machine Experiment. *Nature* 563.7729 (2018): 59.
- [158] Alexander, Larry and Michael Moore, "Deontological Ethics", The Stanford Encyclopedia of Philosophy (Winter 2021 Edition), Edward N. Zalta (ed.), URL = <<https://plato-stanford-edu.tudelft.idm.oclc.org/archives/win2021/entries/ethics-deontological/>>.
- [159] Lee, Minha, Jaisie Sin, Guy Laban, Matthias Kraus, Leigh Clark, Martin Porcheron, Benjamin R. Cowan, Asbjørn Følstad, Cosmin Munteanu, and Heloisa Candello. "Ethics of Conversational User Interfaces." In CHI Conference on Human Factors in Computing Systems Extended Abstracts, pp. 1-7. 2022.
- [160] Prabhumoye, Shrimai, Brendon Boldt, Ruslan Salakhutdinov, and Alan W. Black. "Case study: Deontological ethics in NLP." arXiv preprint arXiv:2010.04658 (2020).
- [161] De Angeli, Antonella. "Ethical implications of verbal disinhibition with conversational agents." *PsychNology Journal* 7, no. 1 (2009).
- [162] Matthias, Andreas. "Robot lies in health care: When is deception morally permissible?." *Kennedy Institute of Ethics Journal* 25, no. 2 (2015): 169-162.
- [163] Seymour, William, Max Van Kleek, Reuben Binns, and Dave Murray-Rust. "Respect as a Lens for the Design of AI Systems." arXiv preprint arXiv:2206.07555 (2022).
- [164] Maureen, Sander-Staudt, "Care Ethics." The Internet Encyclopedia of Philosophy, ISSN 2161-0002, <https://iep.utm.edu/care-ethics/>, access date 29th June, 2022.
- [165] Van Wynsberghe, Aimee. "Designing robots for care: Care centered value-sensitive design." In *Machine Ethics and Robot Ethics*, pp. 185-211. Routledge, 2020.
- [166] Whittemore, Robin, and Kathleen Knafl. "The integrative review: updated methodology." *Journal of advanced nursing* 52, no. 5 (2005): 546-553.
- [167] Snyder, Hannah. "Literature review as a research methodology: An overview and guidelines." *Journal of business research* 104 (2019): 333-339.
- [168] Nag, Procheta, and Özge Nilay Yalçın. "Gender stereotypes in virtual agents." In Proceedings of the 20th ACM International conference on intelligent virtual agents, pp. 1-8. 2020.
- [169] Andreea Niculescu, Dennis Hofs, Betsy Van Dijk, and Anton Nijholt. 2010. How the agent's gender influence users' evaluation of a QA system. In 2010 International Conference on User Science and Engineering (i-USer). IEEE, 16–20
- [170] Annika Silvervarg, Magnus Haake, and Agneta Gulz. 2013. Educational potentials in visually androgynous pedagogical agents. In International Conference on Artificial Intelligence in Education. Springer, 599–602.
- [171] Kevin S Bonham and Melanie I Stefan. 2017. Women are underrepresented in computational biology: An analysis of the scholarly literature in biology, computer science and computational biology. *PLoS computational biology* 13, 10 (2017), e1005134
- [172] Zoe Budrikis. 2020. Growing citation gender gap. *Nature Reviews Physics* 2, 7 (2020), 346–346.
- [173] Neven Caplar, Sandro Tacchella, and Simon Birrer. 2017. Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy* 1, 6 (2017), 1–5.
- [174] Lucía Santamaría and Helena Mihaljević. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* 4 (2018), e156.
- [175] Georgios Halkias and Adamantios Diamantopoulos. 2020. Universal dimensions of individuals' perception: Revisiting the operationalization of warmth and competence with a mixed-method approach. *International Journal of Research in Marketing* 37, 4 (2020), 714–736. <https://doi.org/10.1016/j.ijresmar.2020.02.004>
- [176] Alice H Eagly, Christa Nater, David I Miller, Michèle Kaufmann, and Sabine Sczesny. 2020. Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *American psychologist* 75, 3 (2020), 301.
- [177] Daniel J Rea, Yan Wang, and James E Young. 2015. Check your stereotypes at the door: an analysis of gender typecasts in social human-robot interaction. In International conference on social robotics. Springer, 554–563.
- [178] Brown, Elizabeth, and David I. Perrett. "What gives a face its gender?." *Perception* 22, no. 7 (1993): 829-840.

[179] Burton, A. Mike, Vicki Bruce, and Neal Dench. "What's the difference between men and women? Evidence from facial measurement." *Perception* 22, no. 2 (1993): 153-176.

[180] Baxter, D., McDonnell, M., & McLoughlin, R. (2018, July). Impact of chatbot gender on user’s stereotypical perception and satisfaction. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference* 32 (pp. 1-5).

[181] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. <https://doi.org/10.1191/1478088706qp063oa>

[182] Cuddy, Amy JC, Susan T. Fiske, and Peter Glick. "Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map." *Advances in experimental social psychology* 40 (2008): 61-149.

[183] Rosenberg, Seymour, Carnot Nelson, and P. S. Vivekananthan. "A multidimensional approach to the structure of personality impressions." *Journal of personality and social psychology* 9, no. 4 (1968): 283.

[184] Sutton, Selina Jeanne. "Gender ambiguous, not genderless: Designing gender in voice user interfaces (VUIs) with sensitivity." In *Proceedings of the 2nd conference on conversational user interfaces*, pp. 1-8. 2020.

[185] Petter Bae Bae Brandtzæg, Marita Skjuve, Kim Kristoffer Kristoffer Dysthe, and Asbjørn Følstad. 2021. When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 257, 1–13. <https://doi-org.tudelft.idm.oclc.org/10.1145/3411764.3445318>

[186] Colm Sweeney, Courtney Potts, Edel Ennis, Raymond Bond, Maurice D. Mulvenna, Siobhan O’neill, Martin Malcolm, Lauri Kuosmanen, Catrine Kostenius, Alex Vakaloudis, Gavin Mcconvey, Robin Turkington, David Hanna, Heidi Nieminen, Anna-Kaisa Vartiainen, Alison Robertson, and Michael F. Mctear. 2021. Can Chatbots Help Support a Person's Mental Health? Perceptions and Views from Mental Healthcare Professionals and Experts. *ACM Trans. Comput. Healthcare* 2, 3, Article 25 (July 2021), 15 pages. <https://doi-org.tudelft.idm.oclc.org/10.1145/3453175>

[187] Fitzpatrick, Kathleen Kara, Alison Darcy, and Molly Vierhile. "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial." *JMIR mental health* 4.2 (2017): e7785.

[188] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Paper 415, 1–12. <https://doi-org.tudelft.idm.oclc.org/10.1145/3173574.3173989>

[189] Ta, Vivian, Caroline Griffith, Carolynn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, and Alexia Loggarakis. "User experiences of social support from companion chatbots in everyday contexts: thematic analysis." *Journal of medical Internet research* 22, no. 3 (2020): e16235.

[190] William Seymour and Max Van Kleek. 2021. Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 371 (October 2021), 16 pages. <https://doi-org.tudelft.idm.oclc.org/10.1145/3479515>

[191] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 7–9.

[192] Bryant De'Aira, Jason Borenstein, and Ayanna Howard. "Why should we gender? The effect of robot gendering and occupational stereotypes on human trust and perceived competency." In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pp. 13-21. 2020.

[193] Yen Chiahui, and Ming-Chang Chiang. "Trust me, if you can: a study on the factors that influence consumers’ purchase intention triggered by chatbots based on brain image evidence and self-reported assessments." *Behaviour & Information Technology* 40, no. 11 (2021): 1177-1194.

[194] Cheng Xusen, Xiaoping Zhang, Jason Cohen, and Jian Mou. "Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms." *Information Processing & Management* 59, no. 3 (2022): 102940.

[195] Liew Tze Wei, and Su-Mae Tan. "Exploring the effects of specialist versus generalist embodied virtual agents in a multi-product category

online store." *Telematics and Informatics* 35, no. 1 (2018): 122-135.

[196] Griskevicius Vladas, Joshua M. Tybur, Jill M. Sundie, Robert B. Cialdini, Geoffrey F. Miller, and Douglas T. Kenrick. "Blatant benevolence and conspicuous consumption: when romantic motives elicit strategic costly signals." *Journal of personality and social psychology* 93, no. 1 (2007): 85.

[197] Benbasat, Izak, and Weiquan Wang. "Trust in and adoption of online recommendation agents." *Journal of the association for information systems* 6, no. 3 (2005): 4.

[198] Tay, Benedict, Younbo Jung, and Taezoon Park. "When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction." *Computers in Human Behavior* 38 (2014): 75-84.

[199] Fugate Douglas L., and Joanna Phillips. "Product gender perceptions and antecedents of product gender congruence." *Journal of Consumer Marketing* (2010).

[200] Franke Thomas, Christiane Attig, and Daniel Wessel. "A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale." *International Journal of Human–Computer Interaction* 35, no. 6 (2019): 456-467.

[201] Salanitri Davide, Chrisminder Hare, Simone Borsci, Glyn Lawson, Sarah Sharples, and Brian Waterfield. "Relationship between trust and usability in virtual environments: An ongoing study." In *International Conference on Human-Computer Interaction*, pp. 49-59. Springer, Cham, 2015

[202] Finstad Kraig. "The usability metric for user experience." *Interacting with Computers* 22, no. 5 (2010): 323-327.

[203] Kujala Sari, Ruth Mugge, and Talya Miron-Shatz. "The role of expectations in service evaluation: A longitudinal study of a proximity mobile payment service." *International Journal of Human-Computer Studies* 98 (2017): 51-61.

[204] Jin Yucheng, Li Chen, Wanling Cai, and Pearl Pu. "Key qualities of conversational recommender systems: From users’ perspective." In *Proceedings of the 9th International Conference on Human-Agent Interaction*, pp. 93-102. 2021.

[205] Jin Yucheng, Li Chen, Wanling Cai, and Pearl Pu. "Key qualities of conversational recommender systems: From users’ perspective." In *Proceedings of the 9th International Conference on Human-Agent Interaction*, pp. 93-102. 2021.

[206] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. TickTalkTurk: Conversational crowdsourcing made easy. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW (2020)*, 53–57. <https://doi.org/10.1145/3406865.3418572>

Appendices

- Appendix A : Approved Project Brief
- Appendix B : Published position paper
- Appendix C : Questionnaire Items
- Appendix D : Task Scenarios

Appendix A.

Approved Project Brief

DESIGN
FOR OUR
future

TU Delft

IDE Master Graduation

Project team, Procedural checks and personal Project brief

This document contains the agreements made between student and supervisory team about the student's IDE Master Graduation Project. This document can also include the involvement of an external organisation, however, it does not cover any legal employment relationship that the student and the client (might) agree upon. Next to that, this document facilitates the required procedural checks. In this document:

- The student defines the team, what he/she is going to do/deliver and how that will come about.
- SSC E&SA (Shared Service Center, Education & Student Affairs) reports on the student's registration and study progress.
- IDE's Board of Examiners confirms if the student is allowed to start the Graduation Project.

!

USE ADOBE ACROBAT READER TO OPEN, EDIT AND SAVE THIS DOCUMENT

Download again and reopen in case you tried other software, such as Preview (Mac) or a webbrowser.

STUDENT DATA & MASTER PROGRAMME

Save this form according the format "IDE Master Graduation Project Brief_ familyname_firstname_studentnumber_dd-mm-yyyy". Complete all blue parts of the form and include the approved Project Brief in your Graduation Report as Appendix 1 !

!

family name

Jung

initials

J.Y.

given name

Ji-Youn

student number

5149258

street & no.

zipcode & city

country

phone

email

j.y.jung@student.tudelft.nl

Your master programme (only select the options that apply to you):

IDE master(s):

☐ IPD

☒ Dfl

☐ SPD

2nd non-IDE master:

individual programme:

- -

(give date of approval)

honours programme:

☒ Honours Programme Master

specialisation / annotation:

☐ Medisign

☐ Tech. in Sustainable Design

☐ Entrepreneurship

SUPERVISORY TEAM **

Fill in the required data for the supervisory team members. Please check the instructions on the right !

** chair

Alessandro Bozzon

dept. / section:

SDE / KInD

** mentor

Dave Murray-Rust

dept. / section:

HCD / HICD

2nd mentor

Ujwal Gadiraju

organisation:

Web Information Systems, TU Delft

city:

Delft

country:

Netherlands

comments
(optional)

!

Chair should request the IDE Board of Examiners for approval of a non-IDE mentor, including a motivation letter and c.v..

!

Second mentor only applies in case the assignment is hosted by an external organisation.

!

Ensure a heterogeneous team. In case you wish to include two team members from the same section, please explain why.

IDE TU Delft - E&SA Department /// Graduation project brief & study overview /// 2018-01 v30

Page 1 of 7



Procedural Checks - IDE Master Graduation

APPROVAL PROJECT BRIEF

To be filled in by the chair of the supervisory team.

chair Alessandro Bozzon date 08 - 12 - 2021 signature _____

CHECK STUDY PROGRESS

To be filled in by the SSC E&SA (Shared Service Center, Education & Student Affairs), after approval of the project brief by the Chair. The study progress will be checked for a 2nd time just before the green light meeting.

Master electives no. of EC accumulated in total: _____ EC

Of which, taking the conditional requirements into account, can be part of the exam programme _____ EC

List of electives obtained before the third semester without approval of the BoE

☒ YES all 1st year master courses passed

☐ NO missing 1st year master courses are:

name _____ date _____ signature _____

FORMAL APPROVAL GRADUATION PROJECT

To be filled in by the Board of Examiners of IDE TU Delft. Please check the supervisory team and study the parts of the brief marked **. Next, please assess, (dis)approve and sign this Project Brief, by using the criteria below.

- Does the project fit within the (MSc)-programme of the student (taking into account, if described, the activities done next to the obligatory MSc specific courses)?
- Is the level of the project challenging enough for a MSc IDE graduating student?
- Is the project expected to be doable within 100 working days/20 weeks?
- Does the composition of the supervisory team comply with the regulations and fit the assignment?

Content: ☒ APPROVED ☐ NOT APPROVED

Procedure: ☒ APPROVED ☐ NOT APPROVED

comments

name _____ date _____ signature _____



Personal Project Brief - IDE Master Graduation

The Way You Trust Me

project title

Please state the title of your graduation project (above) and the start date and end date (below). Keep the title compact and simple. Do not use abbreviations. The remainder of this document allows you to define and clarify your graduation project.

start date 08 - 12 - 2021

25 - 05 - 2022

end date

INTRODUCTION **

Please describe, the context of your project, and address the main stakeholders (interests) within this context in a concise yet complete manner. Who are involved, what do they value and how do they currently operate within the given context? What are the main opportunities and limitations you are currently aware of (cultural- and social norms, resources (time, money,...), technology, ...).

"The Way You Trust Me" is a Human-Computer Interaction (HCI) master's thesis project in Industrial Design Engineering (IDE) faculty. This academic project is in collaboration with Web Information System (WIS) at TU Delft. In this project, researchers investigate the process of how users trust a range of conversational agents within a broad domain of Human-AI interaction.

<Background>

Conversational user interfaces (CUIs) have been argued to have advantages over conventional GUIs due to facilitating a more human-like interaction. The rise in popularity of conversational AI agents has enabled humans to interact with machines more naturally. There is a growing familiarity among people with conversational interactions mediated by technology due to the widespread use of mobile devices and messaging services, and this has contributed to a steep rise in the use of conversational agents across several domains.

However, CUIs suffer from abandonment and "disuse" from the user. One way to explain this phenomena is that they are limited to specific speech commands that has been coded for pre-determined domains. Therefore, users' trust toward the system to try different skills of the conversational agent is low. As a result, companies spend a lot of resource on teaching users how to talk to these assistants, such as a book to teach how Siri talks, and sending frequent emails about the new Alexa updates.

<Research Questions>

The term 'conversational agent' refers to various system configurations, such as task-focused customer support chatbots to companion agents like Replika. However, most research in the HCI community investigated different factors that affect trust without considering the unique characteristics each type of conversational agent has. A trajectory of research results revealed that trust in the conversational agent is highly context-dependent, and there are no one-size-fits-all solutions. We suspect this mixed result is derived because of the lack of nuances the current HCI community has yet to implement in their research.

While we acknowledge the effect on different user demographics, we focus on the agent characteristic dimension that ranges from task-focused, smart assistant to companion agent. For example, accuracy could be more critical to a task-focused chatbot, while anthropomorphism is more crucial with companion agents. Therefore, we formulate our first research question as follows:
RQ1. How does user trust unfolds with different types of the conversational agent?

While calibrating trust between automated system and people has been widely investigated in fields such as automated driving, there is a substantial research gap in trust calibration within the domain of conversational agents. As CUIs suffer from abandonment of using the system, we could explain such reaction from the users as "distrust-disuse" in the Trust Calibration graph. We could also hypothesize that the users do "over-trust" in their first impression, due to over-promising advertisement of the system, or due to their lack of mental model towards CUIs powered by NLP technology. Therefore, it is highly relevant research question to investigate how trust calibration works with conversational agents.
RQ2. How does user calibrate their trust with the conversational agent?

space available for images / figures on next page

Personal Project Brief - IDE Master Graduation

introduction (continued): space for images

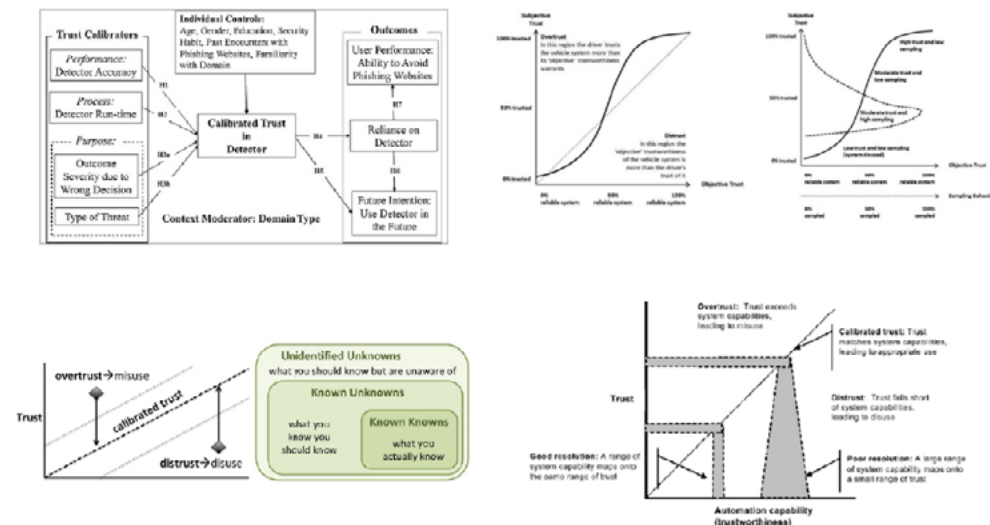


image / figure 1: Different models on how trust works in automated system

Table 1. A taxonomy based on conversation focus.

Type	Focus	Typical sessions	Examples
Virtual companions	Broad, deep	10 to 100's of exchanges	ELIZA, Cleverbot, Tay, Xiaoice, Zo, Hugging Face
Intelligent assistants	Broad, shallow	1-3 exchanges	Siri, Cortana, Alexa, Google Assistant, Bixby
Task-focused chatbots	Narrow, shallow	3-7 exchanges	Dom the Domino's Pizza Bot, customer service bots, Russian trolls, non-player characters

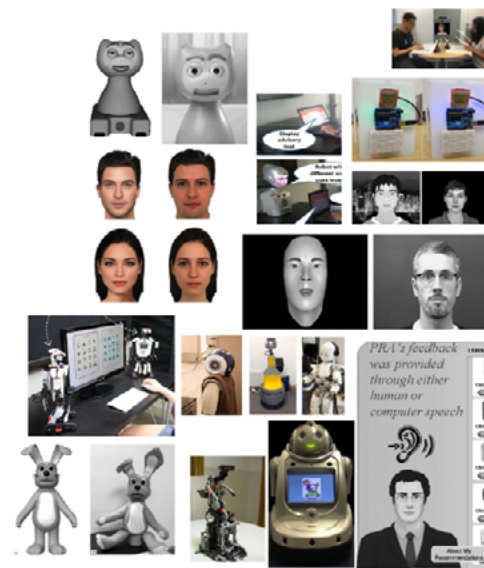


image / figure 2: (Left) Possible taxonomy of CA categories (Right) examples of CAs

Personal Project Brief - IDE Master Graduation

PROBLEM DEFINITION **

Limit and define the scope and solution space of your project to one that is manageable within one Master Graduation Project of 30 EC (= 20 full time weeks or 100 working days) and clearly indicate what issue(s) should be addressed in this project.

In this project, we investigate how trust manifests with conversational agents (CAs) with different functionalities. We investigate with a broader scope of CAs in order to provide high-level contribution on how trust unfolds in the process of usage. Current discourse around trust consists of three main topics. First, researchers can investigate what affects trust. Second, they could see how trust evolves over time. Lastly, researchers could look into what perceivable human behaviours reflect trust. In this project, we focus on the second area of question, that is how trust evolves over time.

In addition, based on the insights we gain, we investigate how we can calibrate trust with CAs. This could be related to the first domain of research mentioned above. However, there is a nuance that we do not just investigate what lowers or enhances the level of trust, but what is the right balance of the level of trust and the actual system capacity. In conclusion, our two main research questions are as follows: RQ1. How does user trust unfolds with different types of the conversational agent? RQ2. How does user calibrate their trust with the conversational agent?

As this project has goals of being academic, we spend less time on exploring and framing the problem space, but instead spend more time on validating and evaluating the research question by spending more time on conducting experiments.

ASSIGNMENT **

State in 2 or 3 sentences what you are going to research, design, create and / or generate, that will solve (part of) the issue(s) pointed out in "problem definition". Then illustrate this assignment by indicating what kind of solution you expect and / or aim to deliver, for instance: a product, a product-service combination, a strategy illustrated through product or product-service combination ideas, In case of a Specialisation and/or Annotation, make sure the assignment reflects this/these.

Deliverables include one research article that communicates our findings on trust manifestation with conversational agents. In addition, design implication should be formulated which should also be addressable in a format of presentation or a video.

As this graduation project aims to conduct about two experiments to validate our research questions, result from these investigation will inform our design implication on trustworthy CUIs. In addition, we will develop a prototype to be used in the experiments.

Personal Project Brief - IDE Master Graduation

PLANNING AND APPROACH **

Include a Gantt Chart (replace the example below - more examples can be found in Manual 2) that shows the different phases of your project, deliverables you have in mind, meetings, and how you plan to spend your time. Please note that all activities should fit within the given net time of 30 EC = 20 full time weeks or 100 working days, and your planning should include a kick-off meeting, mid-term meeting, green light meeting and graduation ceremony. Illustrate your Gantt Chart by, for instance, explaining your approach, and please indicate periods of part-time activities and/or periods of not spending time on your graduation project, if any, for instance because of holidays or parallel activities.

start date **8 - 12 - 2021** **25 - 5 - 2022** end date



This project will be conducted full-time, which means I will work 5 days a week on this project. There are three weeks of holidays planned for this project, making the Graduation presentation date somewhere around the end of May. There is a possibility that this plan could change in the future, due to events that could not be planned in advance such as attending a conference.

There are in total of three experiment planned for this project. At the moment it is highly likely that the Experiment 1 will be qualitative, and Experiment 2 will be quantitative research. Since the schedule is tight, it will be ideal to write the thesis after each phases to prevent a huge workload at the end of the project.

There could be two parallel activities in the beginning of the project: Research elective and the Honours Project. However the plan is to keep my graduation project full-time.

Personal Project Brief - IDE Master Graduation

MOTIVATION AND PERSONAL AMBITIONS

Explain why you set up this project, what competences you want to prove and learn. For example: acquired competences from your MSc programme, the elective semester, extra-curricular activities (etc.) and point out the competences you have yet developed. Optionally, describe which personal learning ambitions you explicitly want to address in this project, on top of the learning objectives of the Graduation Project, such as: in depth knowledge a on specific subject, broadening your competences or experimenting with a specific tool and/or methodology, Stick to no more than five ambitions.

In my undergraduate years at Yonsei University in Korea, my faculty provided quite a unique environment. They advocated a concept called "Integrated Design," where I had to learn four different domains of design: interaction, product, graphic, and even fashion. While exploring such diverse domains, I met the course "User Experience Design" and "Human-Computer Interaction" while pursuing the Cognitive Science minor. It was love at first sight. Something that most intrigued me was how humans and computers speak their language, and I could be the translator and design how their interaction can be. And if I do get the chance to be an HCI designer, how can I be beneficial to society? This project is a part of the journey of myself trying to answer this question.

My interest in the topic of this graduation project started from the C&C course last year (2020). Back then, I explored how a degree of anthropomorphism shown in conversational agents impacted the perceived user trust. This research topic was developed and refined and was investigated in a conversational crowdsourcing agent with different metaphors, each represented in a hierarchical metaphor framework called the Great Chain of Being. This research was produced into a paper and is now getting published in an HCI conference.

So I learned a bit about trust and a bit about the impact of agent representation. While doing so, I found that the field is still filled with many questions with few answers. In my graduation project, I want to push my learnings so far and investigate the area more deeply. Since I aspire to become a researcher in the HCI field, I want to achieve the following competencies in my graduation project.

1. A good understanding of literature

I want to learn the current state of research around trust and conversational user interface (CUI). By doing so, I would like to learn how one can find research gaps and novel research topics. In addition, I want to learn how to formulate clear research questions based on my research interest.

2. Scientific Mindset

I always felt my shortcomings in scientific reporting or researching. I sometimes found myself overstating the outcome of my findings or mixing two different concepts in the field. I want to learn what it means to be scientific and how one can be a scientist that's objective and evidence-based.

3. Technical Skill

In my project, I want to develop technical skills in computer science, especially Natural Language Processing (NLP) or conversational user interfaces engineering. I believe true professionals should have a more hands-on understanding of the topic and tool they're discussing.

FINAL COMMENTS

In case your project brief needs final comments, please add any information you think is relevant.

Appendix B.

Published Position Paper

Gender Choices of Conversational Agent: How Today's Practice Can Shape Tomorrow's Values

Ji-Youn, Jung

Delft University of Technology
Delft, Netherlands
j.y.jung@student.tudelft.nl

Dave Murray-Rust

Delft University of Technology
Delft, Netherlands
d.s.murray-rust@tudelft.nl

Ujwal Gadiraju

Delft University of Technology
Delft, Netherlands
u.k.gadiraju@tudelft.nl

Alessandro Bozzon

Delft University of Technology
Delft, Netherlands
a.bozzon@tudelft.nl

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Copyright held by the owner/author(s).

ACM 978-1-4503-6819-3/20/04.

<https://doi.org/10.1145/3334480.XXXXXXX>

Abstract

When creating conversational agents, designers have to make decisions about the way the agents present themselves. In this position paper, we identify and synthesize ethical dilemmas that conversational interface designers and researchers face around gender of conversational agents. First, we identify three layers that cause tension in designing conversational agents' gender: (i) interactional qualities; (ii) goal-orientation; and (iii) societal issues. We then argue that conversational agent designers and researchers can navigate this problem space by comparing two ethical frameworks: a utilitarian perspective and a dialogical ethics perspective. Finally, we argue that dialogical ethics can be a balanced, ethical lens that can help conversational agent designers and researchers make design decisions about the gender of an agent.

Author Keywords

Gender-inclusive, Dialogical ethics, Design ethics

CCS Concepts

•Social and professional topics → Gender; Please use the 2012 Classifiers and see this link to embed them in the text: https://dl.acm.org/ccs/ccs_flat.cfm

When designing the gender of a conversational agent(CA), several ethical dilemmas rise to the surface. Research find-

ings point us to evidence around how people perceive female agents as friendly and warm, male agents as competent and professional, and gender-neutral agents as creepy [14]. In addition, research has shown that male agents led to higher user trust in the context of airport security [16]. However, several researchers have raised concerns on how designing an agent-based on stereotyped gender roles could reinforce the existing power structures [9, 19, 14]. As technological artifacts and systems acquire embedded values [7] (p.49), it is an ethical design challenge that the conversational user interface (CUI) community holds.

Female gendered agents have a higher chance to experience verbal abuse and sexualization than male-gendered agents or an agent depicted as a nonhuman robot agent [17, 3]. Few studies in education technology have investigated the topic of abusive conversation and reinforcement of cultural stereotypes through gendered CAs [8, 21]. As the natural language processing technology evolves as they learn from the data that agent-user interaction generates, it is of great importance that the system can prevent agents from learning an abusive language. A good example of this danger is a Microsoft Tay, where the agent learned and generated racist, sexist, and anti-Semitic language to the public within 24 hours of its release.

Although studies around gender have a long history, the concept of gender has radically transformed in recent years, leading to a call for reflection. Several recent studies investigated this topic through a contemporary lens. For example, the CHI community has changed their survey system to ask participants' gender in CHI 2016 [18]. Similarly, recent studies have paid attention to practices in image description of web contents [2]. Slumpf et al. [19] presented a conceptual review paper on Gender-Inclusive HCI Research where

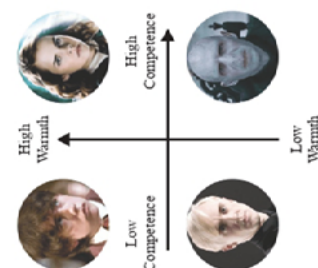


Figure 1: Stereotype Content Model [6], exemplified with famous novel/movie *Harry Potter*'s characters - (from top-left, clockwise) Hermione Granger, Neville Longbottom, Draco Malfoy.

they asserted the need to produce gender-inclusive design, published in 2020. Gender is a complex subject, with a range of political, ethical and social issues attached, especially where devices interact directly with humans using natural language and human-like presentation. The interdisciplinary nature of the CUI community and the current engagement with gender in the design of CUIs makes it a compelling site for developing a robust and caring approach to designing gender.

We present two research gaps that are controversial but are ethical in nature. First, while research has been carried out on the way that users apply gender stereotypes to CAs based on a range of markers [14, 5, 8], little work has been done to calibrate the undesirable outcomes when the anthropomorphized agent does not align with expected stereotypes such as 'female as warm' and 'male as competent'. While this surfaces the implicit assumptions and values embedded in the technology, it only recognizes the danger of reinforcing the stereotype, and does not suggest solutions to overcome the challenge of gender stereotypes.

Second, the consequences of having a non-binary gendered agent have not been investigated deeply. While the concept of fluid gender as a person's identity has recently become an important part of socio-technical debate, most of the HCI community findings that show the effect of a gendered agent only explore the binary male/female concept of gender [19](p.3). The current state of the real-world debate around gender is moving on from this framing, so it is critical that the CA community studies the effect of non-binary gendered agents.

In this position paper, we present our arguments on why these topics matter in the CUI community through an ethical lens, based on the previous research findings. We argue that the CUI community should consider adopting dia-

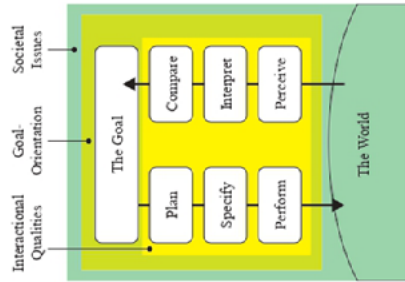


Figure 2: Layers of CA design consideration, augmented on Norman's action cycle [15]

logical ethics in their research practice, which will help develop more robust ethical guidelines for designing gendered agents.

Agents with Gender Stereotypes

Classic works in social perception have shown that people tend to perceive others in category-based information processing automatically. Furthermore, the most salient social categories that help people assign stereotypes are reported to be age, race, and sex [6]. This sensitive categorization develops from an early age, where they also develop their identity through a sense of membership in one gender or others [4]. While such automatic process of categorization make information processing cost-effective, it comes with a cost of stereotyping and discrimination.

As users anthropomorphize CAs [10], previous works have reported how people apply gender stereotypes to conversational agents that shows gender markers [5]. Stereotype content model(SCM), a renowned model in social psychology, explains group stereotypes form along two dimensions of warmth and competency (Figure 1). Considerable amount of findings of user perception towards CAs has been based on this model, where they used warmth and competency as a measure to mediate stereotypical categorization, user expectation, and evaluations [14].

Layers of CUI Design Consideration

Here, several layers can be revealed in considerations for designing gendered conversational agents (Figure 2). First, *interactional qualities* where the point of interaction sparks factors such as enjoyment or engagement. For example, female voices tend to be rated as more likeable [14]. Leading consumer CAs such as Amazon Alexa or Google Assistant is often designed with default feminine voice, and also some widespread consumer reports support user

preference towards female-voiced CAs. Second, the *goal-orientation* level where the characteristics of CAs help the user achieve the final goal of their interaction. For example, users are more likely to choose a male agent to help them solve stereotypically male tasks, such as tasks that require mathematical ability [5]. Finally, a *societal layer* where the danger of reinforcing a gender stereotype exist by deploying gendered agents that conforms to users' — often unintentional — stereotypical expectation [5, 19, 14].

Based on this inspection, CA designers encounter challenging tension between each layer, as illustrated in Figure 2. Existing and widely used axes of warmth and competence help designers provide CA design considerations in 'interactional qualities' and 'goal-oriented' layers. However, few works have explored the consequences and implications of the 'societal' layer, when gendered agents might foster reinforcing gender stereotypes.

Gender neutral agents are unlikeable?

Recently, radical changes in scientific understanding of the endocrine system insisted on biological plasticity. In addition, a traditional understanding of dichotomous gender has changed where a concept of non-binary gender has risen in society. A recent study conducted by Lopatovska et al. [13] reported that nearly half of the participants indicated a preference for a non-binary voice before the experiment. In contrast, gender-ambiguous voices still received the lowest acceptance during the interaction compared to the binary voices. This aligns with previous findings [14] where authors mentioned gender-ambiguous voice creates categorical tension, which leads to unpleasant feelings ("strange, dislikeable, dishonest and unintelligent") towards human users as a possible explanation.

The industry has been developing gender-neutral voices in

recent years. One of the examples is Project Q, a genderless voice assistant to reflect the diversity and reduce the gender bias that got introduced in 2019.¹ Another example is Sam, which got introduced in 2020 and developed by Accenture Labs in collaboration with CereProc.² However, it is early to judge the implication or the consequences of adopting this technology in the real-world.

It is important to notice the tension between people's conscious preference toward non-binary gender agents and unconscious disapproval. Such phenomenon can be explained as a dissonance between people's awareness towards a socially ethical behavior (*societal layer*) and an interactional quality. It will be an assignment for agent designers and researchers to clarify the trade-off between these colliding needs.

Ethics in support of gender choices for conversational agents

Like any ethical dilemmas, different ethical standards can provide moral frameworks to approach the gender choices of CAs. However, we pick two contrasting positions, utilitarian ethics and dialogical ethics.

Utilitarian ethics

Utilitarian ethics will say that the choice which produces the greatest good for the greatest number is the most ethical one. From a Utilitarian point of view, designing an agent that conforms to the gender stereotype seems like a preferable solution to achieve the goal of making agents capable of interacting with users in an efficient and pleasant manner. For example, in the contexts where the credibility of the agent is critical, such as an E-Commerce software where

agents are designed to explain or sell a service or a product, designing an agent with a gender that conforms to the "gender" of its product has shown to improve perceived credibility of the agents(e.g., cosmetics for female agents, hammers for male agents). In addition, perhaps more critically, if there were to be a situation where agents have to direct people for fire evacuation, deploying a male agent that was shown to be more authoritative [20] makes the most sense.

However, if agent designers decide to apply gender stereotypes blindly, they risk strengthening the perception that can result in unfair understanding towards the CA. The game design field has been experiencing a similar dilemma, where gendered game character design can potentially impact youth's understanding of desired gender roles. Moreover, it can shape an individual's body image, self-esteem, self-perception, and expectations from other genders. Some research findings reported over-representation of gender stereotypes in video games [12].

CUI field also holds a similar dilemma to the game design domain, where an overwhelming number of agents today shows feminine markers (e.g., Microsoft's Xiaoice, Amazon's Alexa, Pandorobot's Kuki AI (previously Mitsuku)). As agent abuse and other related challenges call for participation in academic research, we argue that CUI researchers should critically reflect on our research practice that is not solely based on utilitarian values.

Dialogical Ethics

Dialogical ethics is a stance that locates ethics in the communicative ground between people rather than in philosophical thought. Unlike some traditional ethics, dialogical ethics does not censor an opinion or impose any advance restrictions. Instead, dialogue is seen as an epistemological vehicle for learning, where we are doomed to be interdependent

¹<https://www.genderlessvoice.com/>

²<https://youtu.be/mL1n5AEFL4>

in our aim of co-creating moral answers.

Previously, feminist ethicist Koehn [11] suggested dialogical ethics in response to some problems she identified in the traditional feminist ethics, such as care ethics. Adam [1] also recognizes dialogical ethics to provide a more balanced approach to be looked into in her book, where she links feminist ethics with computer ethics.

Instead of giving a single answer, dialogical ethics introduce an ethical way of making a conversation to avoid any wrongdoing in the conversation itself. Surely a conversation could drive one further away from the other if one is to have a self-righteous or a rigid position. Therefore, it is not the verbal exchange of the conversation that matters, but *how* the conversation is made. In conclusion, dialogical ethics focus on fostering mutually acceptable consensus and, therefore, developing a community [11].

According to dialogical ethics, the ethical way of dealing with the dilemma when designing conversational agents would be to openly approach the people who assert the need to develop CAs that go against gender stereotypes. Here, the focus is to have "*the right*" conversation and to prevent any attempt to abstract away from the relevance of the related party's point of view in arriving at a description of the problem at hand [11]. Therefore, an ethical way of solving our dilemma, in the lens of dialogical ethics, could start with developing a guideline to have a respectable conversation around gender and how every party can get an open attitude to acknowledge an insight from each one's point of view.

Final Remarks

As society's gender identity and discourse change, it is natural that the CUI community also faces ethical dilemmas as people anthropomorphize CAs. While many researchers

expressed concerns about adopting gender stereotypes to CA design, there have been no specific guidelines on dealing with such dilemmas with an ethical lens. This position paper shows how ethics can guide the CA designers and researchers to navigate this space. "Just as people bring gender expectations to technology, they can draw gender expectations *from* technology [14]". We believe that the CUI community is one of the important contributors to investigate what could best benefit society.

REFERENCES

- [1] Alison Adam. 2005. *Gender, ethics and information technology*. Springer.
- [2] Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. 2021. "It's Complicated": Negotiating Accessibility and (Mis) Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [3] Sheryl Brahnam and Antonella De Angeli. 2012. Gender affordances of conversational agents. *Interacting with Computers* 24, 3 (2012), 139–153.
- [4] Alice H Eagly, Anne E Beall, and Robert J Sternberg. 2005. *The psychology of gender*. Guilford Press.
- [5] Friederike Eyssel and Frank Hegel. 2012. (s) he's got the look: Gender stereotyping of robots 1. *Journal of Applied Social Psychology* 42, 9 (2012), 2213–2230.
- [6] Susan T Fiske and Steven L Neuberg. 1990. A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology*. Vol. 23. Elsevier, 1–74.

- [7] Luciano Floridi. 2010. *The Cambridge handbook of information and computer ethics*. Cambridge University Press.
- [8] Magnus Haake and Agneta Gulz. 2008. Visual stereotypes and virtual pedagogical agents. *Journal of Educational Technology & Society* 11, 4 (2008), 1–15.
- [9] Charles Hannon. 2016. Gender and status in voice user interfaces. *Interactions* 23, 3 (2016), 34–37.
- [10] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowd-sourcing. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*, April 29-May 5, 2022, New Orleans, LA, USA. ACM New York, NY, USA, 32 pages.
- [11] Daryl Koehn. 2012. *Rethinking feminist ethics: Care, trust and empathy*. Routledge.
- [12] Xeniya Kondrat and others. 2015. Gender and video games: How is female gender generally represented in various genres of video games? *Journal of comparative research in anthropology and sociology* 6, 01 (2015), 171–193.
- [13] Irene Lopatovska, Diedre Brown, and Elena Korshakova. 2022. Contextual Perceptions of Feminine-, Masculine- and Gender-Ambiguous-Sounding Conversational Agents. In *Paper to be published in Proceedings of the 2022 iConference*.
- [14] Clifford Ivar Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge.
- [15] Donald A Norman. 1986. Cognitive engineering. *User centered system design* 31 (1986), 61.

- [16] Jay F Nunamaker, Douglas C Derrick, Aaron C Elkins, Judee K Burgoon, and Mark W Patton. 2011. Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems* 28, 1 (2011), 17–48.
- [17] Annika Silvervarg, Kristin Raukola, Magnus Haake, and Agneta Gulz. 2012. The effect of visual gender on abuse in conversation with ECAs. In *International conference on intelligent virtual agents*. Springer, 153–160.
- [18] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. 2019. How to do better with gender on surveys: a guide for HCI researchers. *Interactions* 26, 4 (2019), 62–65.
- [19] Simone Stumpf, Anicia Peters, Shaowen Bardzell, Margaret Burnett, Daniela Busse, Jessica Cauchard, and Elizabeth Churchill. 2020. Gender-inclusive HCI research and design: A conceptual review. *Foundations and Trends in Human-Computer Interaction* 13, 1 (2020), 1–69.
- [20] Silke ter Stal, Monique Tabak, Harm op den Akker, Tessa Beinema, and Hermie Hermens. 2020. Who do you prefer? The effect of age, gender and role on users' first impressions of embodied conversational agents in eHealth. *International Journal of Human-Computer Interaction* 36, 9 (2020), 881–892.
- [21] George Veletsianos, Cassandra Scharber, and Aaron Doering. 2008. When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with computers* 20, 3 (2008), 292–301.

Appendix C.

Questionnaire Items

Pre-Task Questionnaire

Affinity to Technology:

- Q1. I like to occupy myself in greater detail with technical systems.
- Q2. I like testing the functions of new technical systems.
- Q3. I predominantly deal with technical systems because I have to.
- Q4. When I have a new technical system in front of me, I try it out intensively.
- Q5. I enjoy spending time becoming acquainted with a new technical system.
- Q6. It is enough for me that a technical system works; I don't care how or why.
- Q7. I try to understand how a technical system exactly works.
- Q8. It is enough for me to know the basic functions of a technical system.
- Q9. I try to make full use of the capabilities of a technical system.

Post-Task Questionnaire

Perceived Usability:

- Q1.Using this chatbot is a frustrating experience
- Q2.The chatbot is easy to use
- Q3.I need to spend too much time correcting things with this chatbot.
- Q4.The chatbot meets my requirements

CUI Rapport:

- Q5.The [toothpaste/hair spray/lawn mower] chatbot is warm and caring
- Q6.The [toothpaste/hair spray/lawn mower] chatbot cares about me
- Q7.I like and feel warm toward the [toothpaste/hair spray/lawn mower] chatbot.
- Q8.I feel that I have no connection with the [toothpaste/hair spray/lawn mower] chatbot.
- Q9.The [toothpaste/hair spray/lawn mower] chatbot and I establish a rapport.

Trust-Competence:

- Q10.This chatbot is like a real expert in assessing [toothpaste/hair spray/lawn mower].
- Q11.This chatbot has the expertise to understand my needs and preferences about [toothpaste/hair spray/lawn mower].
- Q12.This chatbot has the ability to understand my needs and preferences about [toothpaste/hair spray/lawn mower].
- Q13.This chatbot has good knowledge about [toothpaste/hair spray/lawn mower].
- Q14.This chatbot considers my needs and all important attributes of [toothpaste/hair spray/lawn mower].

Trust-Benevolence:

- Q15.This chatbot puts my interests first.
- Q16.This chatbot keeps my interests in mind.
- Q17.This chatbot wants to understand my needs and preferences.

Trust-Integrity:

- Q18.This chatbot provides unbiased product recommendations.
- Q19.This chatbot is honest
- Q20.I consider this chatbot to possess integrity.

Desire to cooperate:

- Q21.How likely would you be to cooperate with this chatbot?
- Q22.How likely would you be to help this chatbot?

Intention to adopt (Behavioral Intentions):

- Q23.I will use this [toothpaste/hair spray/lawn mower] chatbot again.
- Q24.I will use this [toothpaste/hair spray/lawn mower] frequently.
- Q25.I will tell my friends about this [toothpaste/hair spray/lawn mower] chatbot.

Appendix D.

Task Scenarios

< Find your perfect [Toothpaste/Hairspray/Lawn Mower]>

You are now going to find the product that best matches the scenario below. To find this, you will interact with the conversational agent who will assist you in your shopping decision. There are a total of 23 toothpaste/hair-spray/lawn mowers, and for each scenario, there’s only ONE product that matches the scenario perfectly. If you submit this matching product, you will be rewarded with a monetary bonus of 0.15 GBP.

Now please start your conversation with the agent on the bottom right corner of this screen. And submit the product that you think fits all the requirements in the scenario above.

Toothpaste

Scenario #1. (Easy)

Imagine your friend is asking for your help to get their toothpaste as the one they’re using right now is almost running out. So they need the toothpaste to be delivered in 1 day. They prefer unscented toothpaste over a fresh scented one. They want their toothbrush to be safe for sensitive teeth.

Scenario #2. (Complex)

Imagine your parents are asking for your help to order toothpaste. They are going overseas by flight to travel for their upcoming vacation. Because of the long flight time, they want to bring their own toothpaste on the plane. Therefore they want travel-sized toothpaste. They want their toothpaste to have a fresh scent. They are looking for a toothpaste that serves total care including gum health. They do not want it to have a whitening function as they heard it is high in its abrasiveness. As their vacation is in a few weeks away, they’re not in a hurry for the delivery time.

Hairspray

Scenario #1. (Easy)

Imagine your friend is asking for your help to get hairspray as they got a new job, and they want to look professional in the office. To have a more natural look, they want it to be either flexible or light hold hairspray. They do not want any textured or shining finish. They prefer it to have aerosol discharge.

Scenario #2. (Complex)

Imagine your parents are asking for your help to order hairspray to use in the wedding they are invited to attend. Because of the long hours, they want it to have a strong or extra strong hold. They want to have a textured finish look. They want the package to arrive within 3 days. They want to have a non-aerosol discharge. Any hairspray for any hair type works for them.

Lawn Mower

Scenario #1. (Easy)

Imagine your friend is asking for your help to get a lawn mower to take care of their lawn. The size of their

lawn is around five thousand square feet, so they want lawn mowers to have 20~22 inches cutting width. They want battery-powered mowers. Their budget is \$345, and the price of a lawn mower should not surpass that amount.

Scenario #2. (Complex)

Imagine your parents are asking for your help to choose a lawn mower for their house. They have around 10,000 square feet of the lawn so they’re looking for lawn mowers with big cutting width, roughly over 27-inch cutting width. They do not care if it’s gas-powered or battery-powered, but they don’t want corded-electric. They want their mower to be rear-wheel drive, not front-wheel drive. Their budget is \$1,700. They want to have three cutting options in their lawn mowers: mulch, bag, and side-discharge.