Gender-Emotion Stereotypes in HRI: Exploring the Role of Gender and Speech Act on the Evaluation of Social Robots

Aafje Kapteijns
5190835

June 27th, 2023

MSc *Artificial Intelligence*
Master thesis
Credits: 44 EC

Utrecht University
Faculty of Science
Department of Information and Computing Sciences
Princetonplein 5
3584 CC Utrecht

Project Supervisor (First Examiner)
Dr. M.M.A. de Graaf

Second Examiner
Dr. E. Herder

Human-Centered Computing Group
Faculty of Science
Utrecht University
Princetonplein 5
3584 CC Utrecht

Abstract

In the field of Human-Robot Interaction (HRI), there has been increased awareness that human-like design of social robots could potentially reproduce societal biases. This study contributes to prior HRI research by conducting a behavioral experiment ($n = 194$), focused on the impact of stereotyping effects concerning a robot's gender and speech act on people's evaluation of warmth, competence and discomfort. The experiment manipulates the Pepper robot's gender (male or female) and the robot's speech act (assertive or affiliative speech) in an online video setting with a between-subjects design. The findings revealed that female robots are ascribed higher competence than male robots, independent of their speech act, and assertive robots are ascribed higher competence than affiliative robots, independent of their gender. Additionally, it was found that participants identifying as female tend to perceive the robot as more competent than those identifying as male. The evaluation of warmth or discomfort do not show significant effects, although there seems to be a tendency for assertive male robots to receive higher discomfort ratings. The results of this study suggest that robot gender as well as user gender and the associated norms and expectations may have complex effects on HRI, mainly in the competence dimension. This highlights the importance of designing social robotics without reinforcing gender bias. Moreover, it emphasizes the need for more theory-driven experiments to address gender issues in HRI, while considering the complexity and diversity of the concept of gender.

Keywords: Human-Robot Interaction · Gendered robotics · Social robotics · Gender stereotypes · Pepper robot

Contents

1 Introduction

In recent developments in the field of Artificial Intelligence (AI) and Human-Robot Interaction (HRI), it has become common practice to equip robots with humanlike characteristics (Bryant et al., 2020). Social robotic design, defined by a humanlike physique and social abilities, is supposed to facilitate interaction and increase aspects of user acceptance (Perugia & Lisy, 2022). Furthermore, it has been shown that when robots are anthropomorphized in embodiment, voice and personality, empathy is increased, which enhances the user experience with the robot (Fink, 2012).

To maximize the utility of robots in terms of acceptability and persuasive effects, an attribute that is often manipulated is the robot's gender representation (Galatolo et al., 2022). However, there is still limited knowledge about what it entails to 'gender' a robot (Perugia & Lisy, 2022). For example, some studies have shown that even without gender being manipulated by the researcher and with the choice for a genderless design, users tend to ascribe gender to robots (Perugia & Lisy, 2022).

Lately, there has been increased awareness in the HRI community that humanlike design invites social categorization and that human form could reproduce gender and racial bias. The implications of AI technologies become evermore visible as the technologies are moving from research labs into mainstream products (West et al., 2019). This can be seen in the widespread usage of voice assistants such as Alexa (Amazon), SIRI (Apple) and Cortana (Microsoft). These technologies all have unmistakable female voices, names and personalities (West et al., 2019). It is argued that these voice assistants are programmed to help and please the user, and are too tolerant of poor treatment. In this way, gendered technological design could reinforce the gender stereotypes that women should be submissive and helpful, and perform lower status occupational roles. The latter is also shown by the fact that many robots performing household chores or providing care, incorporate feminine elements (Marchetti-Bowick, 2009). Therefore, it is crucial to have an understanding about the implications of gendered technological design on society.

It has been demonstrated that robots are also subject to gender stereotypes, when equipped with gender cues (e.g. voice, body shape, hair style, facial cues) (e.g. Tay et al., 2014; Siegel et al., 2009; Eyssel & Hegel, 2012). Namely, Tay et al. (2014) found that people prefer a robot, that performs a task in line with the stereotypes related to that task. Siegel et al. (2009) showed that participants rate the robot from the opposite gender as more credible and trustworthy. Eyssel & Hegel (2012) found that, similar to human psychology, female-gendered robots were perceived as more communal, and male-gendered robots as more agentic.

In human-human interactions it is shown that women engaging in agentic behavior (e.g. dominant, assertive), contrary to gender-stereotypical communal behavior (e.g. helpful, nurturing), often experience backlash effects in the form of social or economic penalties. Namely, it is found that women in higher status occupational roles, possessing high agency, are often perceived as insufficiently communal, and are evaluated as unlikable as a result (Brescoll, 2016; Rudman & Glick, 2001). Studies have shown that men are perceived as more agentic and more competent than women, and

that this could negatively impact women's careers, as agency is associated with successful leadership (Ragins & Winkel, 2011; Rudman & Glick, 2001; Abele, 2003; Brescoll, 2016). Thus, there exists a mismatch between traits typically associated with femininity and traits typically associated with successful leadership, creating the idea that female leaders have to walk a tightrope between being seen as too feminine to be competent, or too masculine to be likable (Williams et al., 2016). Men can be influential without being liked, as long as they are competent, but women have to be likable as well as competent in order to be considered a good leader or colleague (Carli, 2001). These gender-emotion stereotypes are also closely related to language use. Women tend to use language that focuses on building and maintaining relationships with others (i.e. affiliative speech), while men tend to use language that emphasizes their own power and assertiveness (i.e. assertive speech) (Leaper & Ayres, 2007). One example is that when men and women express anger at an equivalent level, the man is still viewed as rational, while the woman is seen as emotionally unstable (Rudman & Glick, 2001).

In this study, I aim to contribute to the existing research within the HRI community with a behavioral experiment, designed to investigate the role of gender, voice and emotion on the user's evaluation of a social robot. The user's evaluation of the robot is assessed in terms of warmth, competence and discomfort, using a standardized scale that is designed to measure the social attributes that people ascribe to robots (Carpinella et al., 2017). Using the knowledge from social psychology research, I aim to research how the gender-emotion stereotypes and implications translate from human-human interaction to HRI. Specifically, I adapt the voice of the Pepper robot (female and male voice) and the speech act (affiliative speech and assertive speech), to find an answer to the following research question: *"How does the gender and speech act of a social robot (Pepper) influence the evaluation of the robot in terms of warmth, competence and discomfort?"*.

The relevance for this project arises from the importance of inclusive gendered technology, in order to avoid the risk of reinforcing gender bias in society. The exploration of gender-emotion stereotypes in HRI and how these impact the evaluation of social robots could provide results that are useful in the debate of inclusive robot design. This study contributes to the broader field of AI, by conducting an experiment within the field of HRI, which develops concepts of interaction between user and robot, bridging the fields of AI, Robotics as well as Social Psychology.

The structure of this study is as follows. First, the research is embedded in related work from the fields of HRI, Gender and Feminist Studies and Social Psychology. Then, the research question and hypotheses derived from the relevant literature are presented. This will be followed by the methodology of this research consisting of two phases. In the first phase the pretests are described, conducted to ensure the effectiveness of the experiment manipulations. In the second phase the setup of the the main experiment is presented. Next, the results obtained from the data analysis are provided. Lastly, the results are discussed and the implications explored in the broader context, after which the findings are concluded.

2 Related work

In this section, I will cover core ideas and terminology from the fields of HRI and Gender Studies, including explanations of the concepts of social robotics and the gendering process. I will briefly touch on the complex field of gender research, as it is crucial to understand the relevance of this topic and the implications of gendered design. Then, I will provide an overview of the implications of gendered robotics on society and examine how gender stereotypes translate to HRI. Furthermore, I will present a variety of experimental studies on gendered robotics, to explore the state-of-the-art knowledge of the impact of a robot's gender on the user's perception and evaluation. Particularly, I will focus on the impact of behavior contradicting stereotypes on the user's evaluation. Lastly, I will focus on related papers exploring the effects of gender stereotypes and the associated challenges in human-human interaction. This related work section forms the basis of my theoretical framework, from which I will derive my hypotheses about the effects of gender-emotion stereotypes in HRI in the next section.

2.1 Social robotics and the gendering process

2.1.1 Anthropomorphizing and social robotics

In the study by de Graaf et al. (2015) a framework of social robotic design is outlined. According to them, social robotic design should focus on interpersonal communication, a lifelike appearance, theory of mind and empathy. It is stated that the aim of social robotics is to create robots that interact with users in a natural and intuitive manner. The goal is for users to be able to relate to the robot and empathize with it.

To accomplish successful interaction and increase the acceptance of the robot, anthropomorphism should be incorporated (Fink, 2012). Anthropomorphism describes the tendency of humans to ascribe humanlike characteristics to non-human objects. In robotic design, this includes humanlike and social qualities, such as the robot's physical shape, behavior and interaction. For example, a robot can be designed with facial expressions, gaze and gestures to give it a humanlike appearance (Fink, 2012).

Thus, anthropomorphic robotic design imitates the human form to achieve higher standards of interaction. The theory of *the uncanny valley*, first formulated by Mori in 1970, states that as robots appear more and more human, people's affinity for them increases, until we fall into the uncanny valley (Mori et al., 2012). The theory describes the reaction of people to a humanlike robot, that closely resembles a human, but fails to hold up the lifelike feeling. According to this theory, people will then quickly shift from empathy to revulsion (Fink, 2012).

Humanlike design does not only shape the way we interact with robots, but also influences how the robot is evaluated (Fink, 2012). In the design process of social robots, researchers make use of methods and concepts from the disciplines of Social Psychology and Communication Science.

Thus, the underlying principle here is that humans prefer to interact with social robots the same way they do with humans (de Graaf et al., 2015).

2.1.2 Gender as a feature for humanlikeness

Gender is one of the strongest informational cues for humans to socially categorize others (Harper & Schoeman, 2003). It is a key part of a humanlike identity, thus, it is often used as the starting point of a robot's persona (Marchetti-Bowick, 2009). In the study by Perugia et al. (2022), it is shown that the degree of humanlikeness is connected to age and gender categorization in robots. Also, Søraa (2017) explores the gendered qualities of current advanced robots and states that the more humanoid a robot becomes, the more gendered it becomes. Perugia & Lisy (2022) present an extensive literature review and suggest that it seems impossible to design humanoid robots without unintentionally prompting gender attribution. Even without deliberately designed gender cues, research has shown that users assign a gender to the robot (Perugia & Lisy, 2022). For example, in the research of Jackson et al. (2020), it was found that even with a genderless voice, other signifiers such as role and speech act can lead to implicit gendering. These studies highlight that there is still little knowledge about what it entails to gender a robot and the implications of such design choices.

2.1.3 The concept of gender

Following the example of Perugia & Lisy (2022), I will adopt the viewpoint that the gendering of a robot is a two-step process of *gender encoding* (gender as attributed by the designers) and *gender decoding* (gender as attributed by the users). Furthermore, the *genderedness* of a robot is the property of being gendered as a result of this process (Perugia & Lisy, 2022). According to them, Gender Studies and Feminist theories can enrich the understanding of what gendering of a robot entails. It is therefore crucial to be aware of the evolution and implications of the concept of gender, in order to comprehend how gendered technological design can have an effect on society (Perugia & Lisy, 2022). Thus, in this paragraph I will briefly touch upon the concept of gender, before presenting the implications of gendered technological design and robotics. I will reference some important works in the field of Gender Studies, as it is necessary to understand further related work and the relevance of this thesis. However, due to the complexity of this research space, it is by no means a complete definition of the concept of gender, as this falls out of the scope of this thesis.

In the work of Lorber (2018) gender is viewed as a social institution and one of the most important concepts through which human beings organize their life. Gender as a social construct is not equivalent to sex, and does not automatically follow from the reproductive organs, the main difference between females and males (Lorber, 2018). Once gender is ascribed, individuals are subject to gendered norms and expectations, which directs how women and men should act. Lorber (2018) states that men and women are perceived as different, even if they are doing the same tasks.

Same tasks are sometimes given different job titles for men and women, such as executive secretary and administrator assistant.

Gender is often understood on a binary scale, either male (masculine) or female (feminine) (Richards et al., 2016). At birth, sex is linked to gender, to gender roles and to sexuality. However, across time and cultures non-binary, or genderqueer identities have been present (Richards et al., 2016). With the emergence of feminist and queer theory the binarism of gender has been contested (Perugia & Lisy, 2022).

Thus, gender is a complex social construct that goes beyond biological differences between male and female. Gendered norms and expectations can limit the opportunities of individuals or restrict their behavior. As HRI is a human-centred field of study, researchers strive to account for the complexity of the research space. Here, acknowledging the existence of non-binary gender identities helps to promote inclusivity in both research and design. For instance, Winkle et al. (2023) demonstrate ways to bring feminist principles to the field of HRI and highlight the importance of researching gender. They call on researchers to be attentive of power structures, based on concepts such as gender, race, class, religion, sexuality and more, in their research question, approach and data collection.

2.1.4 Summary

In sum, the aim of social robotics is to create robots that are able to interact with users in a natural manner. Social robotic design often incorporates humanlike features to enhance interaction and increase acceptance. As robots become more social and humanlike, they also become more gendered. Studies have suggested that even without deliberate gender cues, users tend to assign gender to a robot, highlighting the fact that there is still little knowledge about what it entails to gender a robot and what the implications of gendered design are. This makes it crucial to research gender within the field of HRI, while acknowledging the complex nature of gender as a construct involving norms, expectations and power dynamics. In order to promote inclusivity, researchers should be attentive to non-binary gender identities and the impact of power structures in their research and design approaches. The next section will be dedicated to the implications of incorporating humanlike (and thus gendered) features in robotic design.

2.2 The implications of gendered robotic design

2.2.1 Gender stereotypes in robotics

An increasing concern among researchers in HRI is that the design of humanlike robots may lead to the reproduction of social categorization and biases, based on the robot's appearance (e.g. Perugia & Lisy, 2022; Alesich & Rigby, 2017). Studies have shown that most gendered robots currently being developed are female, and are often designed based on preconceived notions of gender roles and physical characteristics (Alesich & Rigby, 2017). Particularly female-gendered robots are designed

with hyper-feminine physical characteristics and often carry out tasks traditionally associated with women (Alesich & Rigby, 2017).

People tend to socially categorize robots based on physical characteristics, such as gender and racial cues. Social categorization is defined as as a cognitive process, by which individuals group people based on shared characteristics, that occurs to simplify and systematize perceptive information. This stems from the *Social Identity Theory*, first introduced by Tajfel (1979), that suggests that people have a natural tendency to categorize themselves and others in groups, with a shared sense of identity and social values. This can lead to incorrect perceptions, the stimulation of differences between groups and the over-generalisation of an individual based on the group they belong to. This process is called stereotyping (Taylor, 1981).

It has been demonstrated that robots are also subject to gender stereotypes, when designed with facial cues, such as hair style, or body shape (e.g. Eyssel & Hegel, 2012; Bernotat et al., 2017; Abele, 2003; Chita-Tegmark et al., 2019; Tay et al., 2014). These studies found that tasks perceived as stereotypically male were considered more suitable for male robots than female robots and vice versa. Furthermore, the female-gendered robots were perceived as more communal, and male-gendered robots as more agentic, both stereotypically ascribed traits of respectively women and men. Chita-Tegmark et al. (2019) evaluated the impact of gender on the perceived emotional intelligence of robots and humans and compared the outcomes. They showed that in an office scenario, by only altering the gender with minimal markers, the emotional intelligence perceptions are affected for robots the same way as for humans. These studies show that gender stereotypes can be transferred from human-human interactions to human-robot interactions.

2.2.2 Gendered technologies in society

In a recent UNESCO report, titled *I'd Blush if I could*, the proliferation of female digital voice assistants is examined, to highlight the ethical implications of gendered AI (West et al., 2019). It is stated that most voice assistants have a female voice and name (i.e. Amazon with Alexa, Apple with SIRI, Microsoft with Cortana). The report argues that the voice assistants are programmed to be "obliging, docile and eager-to-please helpers" and that they are too tolerant of poor treatment. The latter is also reflected in the title of the report, which corresponds to the answer SIRI gave to an inappropriate remark or insult[1].

The way these voice assistants are designed can reinforce harmful stereotypes that women should be submissive, serving and tolerate mistreatment (Winkle et al., 2021). This is becoming especially relevant now that technologies are moving from research labs to mainstream consumer products (West et al., 2019). Robots that are created to perform jobs for consumers, are often designed with gender-specific characteristics. For instance, many robots that perform household chores or

[1]SIRI, meaning 'beautiful woman who leads you to victory' in Norse, is the voice assistant created by Apple. The answer it gives to an insult has been updated to the neutral answer "I don't know how to respond to that.". Companies such as Amazon and Apple justify their decision of using a female voice as the default (and for the first years as sole option) by citing work demonstrating that people prefer a female voice to a male voice.

provide care, incorporate feminine elements. They embody idealized versions of women, from "perfect" physiques to traditional gender roles (Marchetti-Bowick, 2009). In the study by Hover et al. (2021) online reactions to video's featuring gendered, humanlike robots were analyzed. It was found that humanlike, female robots were the most likely to evoke negative attitudes and be subject to sexualization and sexism.

2.2.3 Summary

HRI researchers have found that the design of humanlike robots may lead to social categorization and biases, based on the robot's appearance. Studies have shown that certain gender stereotypes can be transferred from human-human interaction to human-robot interaction. Despite these facts, it has been stated that gendered robots are predominantly female, designed with hyper-feminine physical characteristics and often carry out tasks associated with traditional gender roles. Gendered technology is increasingly used in mainstream consumer products, such as female digital voice assistants or female embodied robots performing household chores among others. These products are often designed in ways that perpetuate harmful gender stereotypes. This highlights the importance of carefully considering the ethical implications of gendered technologies. In the next section, I will present experimental studies that explore the impact of a robot's gender on user perception, and show how gender stereotyping and its implications can be directly seen in the robot's evaluation.

2.3 Evaluations of gendered robotics

2.3.1 Gender-task experiments

Several researchers have studied the impact of a robot's gender on people's perception of it, including measures such as trust, discomfort, engagement, likability or emotional intelligence (e.g. Chita-Tegmark et al., 2019; Bryant et al., 2020; Neuteboom & de Graaf, 2021; Reich-Stiebert & Eyssel, 2017). Some HRI studies specifically focus on the relation between a robot's gender and the assigned task. Namely, Tay et al. (2014) found that people prefer a robot that performs a task which is in line with existing gender stereotypes of that respective task. Also, in the paper by Perugia & Lisy (2022), it is stated that robots with feminine gender cues are often perceived as the better fit to perform stereotypically female tasks, such as providing services in domestic settings. On the other hand, Reich-Stiebert & Eyssel (2017) have found that when a mismatch occurred between robot gender and the gender typically associated with the task, people were more willing to engage in learning processes of academical tasks with the robot. According to this paper, it would be possible to challenge persisting gender stereotypes and therefore mitigate gender bias in this context.

2.3.2 Impact of norm breaking behavior on evaluation

The gender-task mismatch is one form of norm breaking behavior, as the incongruity between the robot's gender and the gender typicality of the task results in a deviation of the user's expectations of

the robot's performance. Therefore, the mismatch impacts the evaluation of the robot. Neuteboom & de Graaf (2021) found that a gender-task mismatch could lead to dehumanization of the robot into an emotionless object. Namely, it seemed that dehumanization occurred exclusively with female robots performing an analytical task and male robots performing a social task. Carpinella et al. (2017) found that the humanlikeness of a robot influences the level of discomfort in the user's evaluation. The more machinelike a robot appeared, the higher the level of discomfort it evoked. Therefore, a mismatch between gender and stereotypical behavior, resulting in dehumanization, could evoke a higher level of discomfort in users. Coming back to the theory by Mori et al. (2012) of the Uncanny Valley, the negative "uncanny" feeling can be triggered by the appearance of the robot, but also by the behavior and whether or not it is in line with stereotypes.

Another form of norm breaking behavior is researched by Jackson et al. (2020). In their paper, the effects of robot and human gender in noncompliance interactions are studied and they found that it is more favorable for male robots to reject commands from the user than for female robots. Winkle et al. (2022) also look at the robot's responses in interactions, as they investigate the impact of different responses to sexist abuse on the credibility of the robot. They found that the robot's credibility was significantly higher when it presented a norm breaking counterargument to the abusive statement, as opposed to refusing engagement or launching an attack. According to them, these responses can boost the credibility of the robot, while also avoiding the propagation of gender stereotypes.

2.3.3 Impact of user gender on evaluation

Jackson et al. (2020) findings also suggest that a robot is evaluated more positively if their gender matches that of the user. Therefore, it is important to take the gender of the user into account, when studying robot gender. The study by Siegel et al. (2009) stresses this point as well, although with a different conclusion. They researched the impact of gender on the persuasion of a humanoid robot, by letting the participants respond to a donation request of the robot. It was found that men were more likely to donate money to the female-gendered robot, while women did not seem to have a preference. Furthermore, the participants generally rated the robot of the opposite gender as more trustworthy, credible and engaging (Siegel et al., 2009). Also, in the research by Otterbacher & Talias (2017), it was found that other-gender participants were more susceptible to gender stereotypes than same-gender participants. When female robots appeared more communal, and when male robots appeared more agentic (in line with the respective stereotypes), an uncanny reaction was more likely to occur in other-gender participants than in same-gender participants. Lastly, Galatolo et al. (2022) also replicate the findings by Jackson et al. (2020) regarding the complex interaction between robot, observer and interactant gender. They conducted a video-based study featuring two actors and a robot, that challenged gender stereotypes about men and women. They found that women were more impressed than men, when the robot challenged a stereotype about men. These studies demonstrate how gendered expectations and robot evaluations

are influenced by various factors, such as robot gender, participant gender and the experiment context.

2.3.4 Summary

HRI researchers have studied the influence of a robot's gender on the user's evaluation. Studies have shown that people tend to prefer robots that perform tasks in line with existing gender stereotypes. It has been found that the occurrence of a gender-task mismatch, or another form of behavior that challenges stereotypes, could result in a deviation of the user's expectation of the robot's performance, impacting its evaluation and potentially leading to dehumanization or even increased discomfort. However, other studies have found that the uttering of norm breaking arguments by robots could boost credibility, and that the gender-task mismatch could help users engage in learning processes. Furthermore, it has been shown that the user's gender greatly impacts the evaluation of a gendered robot, although there have been conflicting findings. I believe that these conflicting findings emphasize the relevance of researching the impact of gender, both robot and user gender, and norm breaking behavior on robot evaluations. In the next section, I will explore gender stereotypes and associated implications of norm breaking behavior in human-human interaction.

2.4 Gender and emotion in psychology

2.4.1 The interplay of gender and emotion

In human psychology, one of the strongest stereotypes held in Western culture is the belief that women are more emotional than men (Brescoll, 2016). Namely, people believe that women express more emotions, except for anger and pride, which are emotions typical for men (Brescoll, 2016). Furthermore, when women express anger, this is more negatively evaluated than when men express anger. This gender-emotion stereotype leads to a mismatch between traits typically associated with femininity and those considered necessary for success and influence (Ragins & Winkel, 2011).

Eagly & Karau (2002) have proposed a theory that explains the incongruity between the female gender role and a leadership role, known as *the role congruity theory of prejudice towards female leaders*. This theory states that this mismatch leads to two types of bias, namely, women are viewed less favorably as potential leaders than men, and when women behave in the same manner as a male leader, it is evaluated less positively. This leads to a greater difficulty for women to become successful leaders.

The prejudice towards female leaders is further explored in a theoretical account by Ragins & Winkel (2011). They investigate the impact of gender roles on the display and evaluation of emotion in work relationships. It is stated that men are seen as more agentic and more influential than women, and that women are seen as less competent than men. Abele (2003) states that if women are seen as less agentic, this can negatively impact their careers, as agency (i.e. dominance,

assertiveness, aggression) can be crucial for a successful performance as a leader. On the other hand, when men are seen as less communal, this can harm their personal relationships, as communality (i.e. nurturing behavior, kindness) can be important for successful relationships. Carli (2001) studied the stereotypes concerning communality and agency, and attributed them to the traditional societal gender roles, with women more often in domestic, lower status occupational roles and men more often in higher status occupational roles. She states that this relates to the societal expectations of men to behave more agentically and women to behave more communally.

Hentschel et al. (2019) researched the complex dimensions of communality and agency and their impact on how men and women perceive themselves and others. They found that men tend to perceive women as less agentic than men, and less agentic than women described themselves. The agency dimensions were leadership competence, assertiveness and independence. Considering these key agency qualities, women viewed themselves as less competent and less assertive leaders than men viewed themselves. Moreover, women described other women as less assertive than men. However, contrary to men's perceptions, they described other women as equally leadership competent as men. Thus, men perceive women as less leadership competent than women perceive women.

2.4.2 Competence and likability challenges for female leaders

Women who do engage in agentic behavior, contrary to feminine stereotypes, often experience backlash effects in the form of social or economic penalties. Rudman & Glick (2001) and Brescoll (2016) have examined these backlash effects for women in leadership positions, and state that because these women are perceived as insufficiently communal they are evaluated as unlikable. Moss-Racusin et al. (2010) state that when men behave insufficiently agentic, they are also at risk of backlash effects, mostly in terms of likability. The researchers propose that this backlash occurs because of a perceived incongruity between men's behavior and societal expectations of masculinity. Rudman & Glick (2001) refer to these different standards of men and women's emotional behavior in the workplace as an "emotional double bind". Studies have supported this concept, demonstrating for example that when a woman expresses anger at the same degree as a man, only she is viewed as emotionally unstable (Rudman & Glick, 2001).

The study of Carli (2001) concerning the impact of gender to social influence adds to this, because she states that a man can be influential without being liked as long as he is competent, while a woman must be competent and likable simultaneously, in order to be considered a credible leader. Williams et al. (2016) have researched this phenomenon within STEM (science, technology, engineering and math). They describe mechanisms of gender bias in everyday workplace scenario's, such as *the Tightrope* and *Prove-it-again* patterns. The first entails that women have to walk a tightrope between being seen as too feminine to be competent, or too masculine to be likable. The latter involves women having to provide more evidence of their competence than men.

Another commonly researched concept related to gender bias in the workplace, is *the queen bee phenomenon* (Williams et al., 2016; Derks et al., 2016). This refers to women who distance

themselves from other women and adjust to the culture in male-dominated work environments, in order to achieve individual success. This is often driven by the fact that women are also biased against other women in such settings, as well as by the limited number of opportunities for women within these workplaces. Over 50% of the STEM scientists that were interviewed by Williams et al. (2016) reported experiencing this phenomenon in their own workplace.

2.4.3 Assertive and affiliative speech

Gender and emotion are closely related to language, as researchers have found differences in typically male and female language use (Leaper & Ayres, 2007; Park et al., 2016). For example, women are more likely than men to use language to create and maintain connections with others (i.e. affiliative speech), and men are more likely than women to use language to assert dominance (i.e. assertive speech) (Leaper & Ayres, 2007). A meta-analytic review by Leaper & Ayres (2007) examined gender variations in adults' language use, in which they make the distinction in assertive and affiliative speech acts. The assertive speech act involves the promotion of one's personal agency, while the affiliative speech act is more related to the communality dimension, and functions to engage with others. Examples of assertive speech are directive or task-oriented statements, imperative statements, suggestions or critical remarks, while examples of affiliative speech include supportive and collaborative statements, active understanding and acknowledgement of the other person's contributions. Thus, these findings support the existence of gender-emotion stereotypes concerning the agency and communality dimensions and show that they are closely related to language use.

2.4.4 Summary

In sum, one of the strongest gender stereotypes is that women are seen as being more emotional than men. These gender-emotion stereotypes can also be seen in language use, as women tend to use affiliative speech to maintain connections, while men tend to use assertive speech to assert dominance. Thus, there exist different expectations for how men and women should behave and communicate in society. It is found that there exists a mismatch between traits typically associated with femininity and traits associated with leadership, resulting in women facing greater challenges to become successful leaders. Women engaging in agentic behavior often experience backlash in the form of social or economical penalties and have to provide more evidence of their competence than men. Men who deviate from agentic behavior also experience backlash, mainly in the form of likability. Furthermore, there exists a difference in how men and women evaluate others, as men describe women as less competent than women describe women. These studies from the fields of Social Psychology and Communication Science underscore that norm breaking behavior in humans, by challenging gender-emotion stereotypes, can have great impact on evaluations by others in terms of likability and competence. In the next section, my research question and hypotheses will be presented, in which I focus on whether or not these findings translate to HRI.

3 Research question and hypotheses

This study presents a behavioral experiment, designed to investigate the impact of gender, voice and emotion on the evaluation of a social robot. The aim of this thesis is to research how the gender-emotion stereotypes of agency and communality and implications such as the backlash effects translate from human-human interaction to human-robot interaction. Specifically, as a gender-emotion stereotype, the distinction made by Leaper & Ayres (2007) regarding gender variations in language use is used, namely, affiliative speech and assertive speech. The evaluation of the social robot by the user is based on the warmth, competence and discomfort dimensions, using a standardized scale that is designed to measure the social attributes that people ascribe to robots (Carpinella et al., 2017).

In the experiment, the voice (male or female) and the speech act (assertive or affiliative speech) of the Pepper robot is adapted, in order to find an answer to the following research question: *"How does the gender and speech act of a social robot (Pepper) influence the evaluation of the robot in terms of warmth, competence and discomfort?"*.

First, I hypothesize that the backlash effects as found in human-human interactions, following the findings of Rudman & Glick (2001) and Brescoll (2016), could be found in the warmth dimension of the user's evaluation of the female robot engaging in behavior contradicting gender-emotion stereotypes. When a female-voiced robot speaks in an assertive way, contrary to the gender-emotion stereotype of speaking in an affiliative way, I hypothesize that the user will give the robot a likability penalty, indicated by lower warmth ratings than the male-voiced assertive robot. Thus, the first hypothesis is as follows:

H1: "*The female-voiced assertive Pepper robot will receive lower scores in the warmth dimension than the male-voiced assertive Pepper robot.*"

Second, I follow the findings of Ragins & Winkel (2011) that stereotypical feminine traits mismatch with traits of successful leaders (e.g. competence). According to them, this also results in the fact that women are viewed as less competent than men. Therefore, I hypothesize that the female-voiced Pepper robots will be perceived as less competent than the male-voiced Pepper robots, independent of the robot's speech act. Thus, the second hypothesis is as follows:

H2: "*The female-voiced Pepper robots will receive lower scores in the competence dimension than the male-voiced Pepper robots, independent of speech act.*"

Third, as emphasized by Jackson et al. (2020), I will take the participant's gender into account. Here, I follow the findings of Hentschel et al. (2019), stating that men generally describe women as less competent than women describe women. I hypothesize that this phenomenon also has an impact on HRI, and that participants that identify as male will perceive the female-voiced Pepper

robot as less competent in comparison to participants that identify as female, who will display this to a lesser extent. Thus, the third hypothesis is as follows:

H3: "*Participants that identify as male will give lower scores in the competence dimension to the female-voiced Pepper robots than participants that identify as female.*"

Lastly, I follow the findings of Neuteboom & de Graaf (2021) that when a mismatch between the gender of the robot and stereotypical behavior occurs, this could lead to dehumanization. As stated by Carpinella et al. (2017) and in line with the "uncanny valley theory" by Mori et al. (2012) , the less humanlike a robot, the more discomfort it evokes. Therefore, I hypothesize that the robots engaging in norm breaking behavior, contradicting gender-emotion stereotypes (the female-voiced assertive robot and the male-voiced affiliative robot), will receive higher scores in the discomfort dimension. Thus, the fourth hypothesis is as follows:

H4: "*The female-voiced assertive Pepper robot and the male-voiced affiliative robot will receive higher scores in the discomfort dimension, than the female-voiced affiliative Pepper robot and the male-voiced assertive Pepper robot.*"

4 Methodology

In this study an online video experiment is conducted manipulating the robot's gender and the speech act used by the social robot to investigate the impact of gender and speech act on the evaluation of the robot in terms of warmth, competence and discomfort. This creates a 2 (robot gender: male and female) x 2 (speech act: assertive speech and affiliative speech) between-subjects design. Participants watched one of the recorded videos featuring the Pepper robot and answer several questions in an online questionnaire. This study was approved by the Ethics Committee of Utrecht University.

4.1 A note on positionality and inclusive HRI research

Winkle et al. (2023) present principles for incorporating a feminist approach in HRI research, aiming to support a more ethical HRI practice. According to them, feminist HRI implies considering not only the robot's physical design, but also the context that the robots are used in, and how the people around them are positioning the robots in practice. Moreover, they state that an inclusive and feminist perspective requires researchers being transparent and aware about their positionality and point of view. A researcher's personal and professional experiences, political and ideological stances and demographics, such as gender, age and race, form their positionality (Soedirgo & Glas, 2020).

As a female researcher, I am personally affected by gender-stereotypical biases in a Western, metropolitan context. Therefore, I am biased in my understanding of the concept of gender, which results in a biased point of view in my approach, methods and the analysis of the data. As a feminist researcher, I am personally committed to the topic of gender-emotion stereotypes. Because of my feminist viewpoint, I had a drive to focus on stereotypes harmful for women, leading me to the topic of likability and competence backlashes for female leadership and assertiveness. It is important to note that I acknowledge that the binary gender concept is limited. However, in this specific topic, the drawbacks for women are more visible than for men, regarding career success, sociability and influence (Ragins & Winkel, 2011). This highlights the relevance to study the evaluation of men versus women, unfortunately limiting the view on gender to a binary construct. Moreover, I acknowledge that men are also harmed by these gender-emotion stereotypes, but due to the limited scope of this thesis, I decided not to include this in my study.

4.2 Robotic platform

The robotic platform used in this study is the Pepper robot, developed by Softbank Robotics (SoftBank Robotics America Inc., 2023) (see Figure 1). The Pepper robot is a humanoid social robot, capable of interacting with humans through speech, gestures and expressions. The robot is 1,22 meters tall and equipped with various features, including touch sensors, an integrated tablet, cameras and a microphone. In this study, Pepper was programmed using the following software: the

Docker platform, the Social Interaction Cloud (SIC) framework, Google Dialogflow and a custom Python SIC application. Docker is a containerization technology that allows developers to access and deploy applications that utilize the Dialogflow services (Merkel, 2014). The SIC framework, accessed using Docker, enables the connection with the Pepper robot (Koeman et al., 2022). Google Dialogflow, a cloud-based platform to create conversational applications, was used to access the Cloud text-to-speech API on the Pepper robot (Google Cloud Text-to-Speech API, 2023).

The Pepper robot is designed without explicit gender cues (Pandey & Gelin, 2018). However, researchers have challenged the claim that the Pepper robot's shape is gender neutral. For example, Jackson et al. (2020) did not use the Pepper robot in their experiment, because "[...] we believe its morphology is implicitly feminine, with a narrow waist, wide hip joint, and a skirt-like shape to the lower half." To eliminate the influence of confounding factors related to the gender manipulation in this experiment, I specifically displayed only the upper half of the Pepper robot, excluding the waist and hip area. Figure 1 shows the Pepper robot, as depicted in the experiment videos. The default color settings of the robot's lights (blue ear lights and white-pink eye lights) and the default interface of the integrated tablet were used.

In this study, video recordings are made of the Pepper robot. The reason for this design choice, is that it is found that video recordings are one of the most reliable methods for measuring behavior and interaction, as the researcher can make repeated viewings as well as nonverbal context available to the participant (Leaper & Ayres, 2007). Furthermore, with a video recording instead of live interaction, the researcher can ensure that each participant experiences the exact same interaction with the robot.

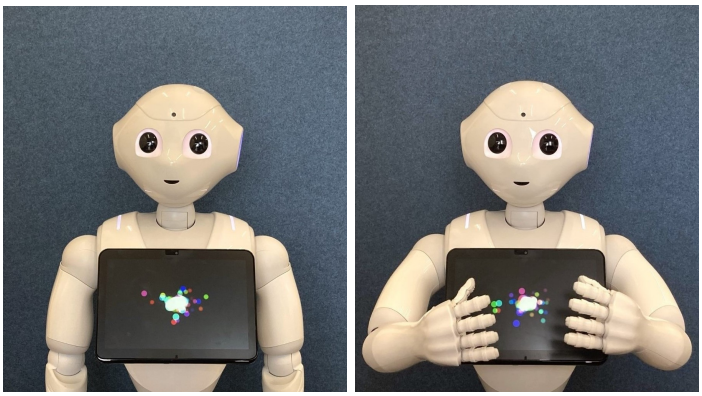


Figure 1: Images of the Pepper robot as depicted in the experiment videos. The Pepper robot uses hand gestures while speaking.

4.3 Phase 1: Pretesting experiment manipulation

Two pretests were conducted to ensure the effectiveness of the manipulations central to the research design of the main experiment of this study. These pretests served as validation measures to confirm that the manipulations were perceived correctly by the participants. Specifically, two separate pretests were conducted: one to test the gender manipulation (male and female) and one to test the speech act manipulation (assertive speech and affiliative speech). The following sections provide details of the two pretests. The results of the pretests served as the basis for the choices made regarding the gender and speech act manipulation of the Pepper robot in the main experiment.

4.3.1 Pretest 1: Gender manipulation

Voice as gender cue: The Pepper robot's gender is manipulated by changing its voice. Perugia et al. (2021) state that vocal cues are more powerful than facial cues to attribute gender to a robot. Furthermore, they state that multiple gender cues might layer and affect each other, leading to my decision to only use voice as a gender cue in my experiment.

Bryant et al. (2020) pretested various pitch settings of computer-generated voices in their experiment with the Pepper robot. Additionally, a gender-neutral voice is used in their experiment. This study is limited to a male and a female voice. This is supported by the fact that a genderless voice is often not perceived as such (Jackson et al., 2020; Galatolo et al., 2022; Perugia & Lisy, 2022). Therefore, employing a genderless voice as a vocal gender cue does not automatically result in the perception of a gender-neutral robot. For this reason, the current study is limited to a binary robot gender representation.

The Pepper robot's default voice is supposed to be childlike and androgynous, with the possibility of varying the pitch (Pandey & Gelin, 2018). In the first pretest, the default voice of the Pepper robot is manipulated in pitch from setting 0 (the minimum level) to 100 in steps of 20 (creating six different voices), to test whether this computerlike voice default for the Pepper robot platform, can be used as gender cue in the main experiment. A pitch higher than 100 was not chosen to be pretested, as it was unpleasantly high.

Participants and Procedure: The participants of the first pretest ($n = 20$) were recruited through convenience sampling. They were asked to evaluate the six voices on a 7-point Likert scale from *Male* to *Female*, with an explicitly stated neutral option (the middle scale option). The evaluation aimed to assess the participants' perception of the voices, which were presented to them as audio files in a random order. To create the audio files, the default voice of the Pepper robot was recorded, uttering the neutral statement (adapted from Kuchenbrandt et al. (2014)): *"According to my watch, it is now a quarter to three. The train will leave in five minutes."*.

Results: The results of the pretest are shown in Appendix A. The voices with pitch setting 80 and 100 were unanimously perceived as female by all participants, indicated by their responses varying between scale options 5, 6 and 7 on the 7-Point Likert scale. However, not one of the voices was unanimously perceived as male. The voice with pitch setting 0 received the most responses

indicating male, with 12 out of 20 responses ranging between scale options 1, 2 and 3. 8 participants perceived this voice as neutral or female.

The Kruskal-Wallis statistical test was conducted on the six voices, and with a post hoc analysis the voice with pitch settings 0 and 100 were compared. The voice with a pitch setting of 0 was perceived as more male than the voice with a pitch setting of 100 ($p < .001$). However, due to the ambiguity in the participants responses for the voice with pitch setting 0, relying on this voice as the male gender cue could be problematic. It is unclear whether the gender manipulation is effective, as the pretest did not provide a consistent indication of male gender perception.

Conclusion: In conclusion, the results of the first pretest indicate that the default voice of the Pepper robot with pitch settings 80 and 100 were perceived as female. However, there was no pitch setting within the Pepper robot's system that was unambiguously perceived as male, as even the voice with pitch setting 0 had a large variety in the responses. To ensure that the gender manipulation is perceived correctly by the participants in the main experiment, the choice was made to use the Google Cloud text-to-speech API supported humanlike voices. Choosing humanlike text-to-speech software instead of default computerlike voices is common in the HRI community. For example, Chita-Tegmark et al. (2019) and Reich-Stiebert & Eyssel (2017) used the Mac OS text-to-speech voices in their experiments. Moreover, the research by McGinn & Torre (2019) stated that the default voice of the Pepper robot was rarely matched with its appearance by participants, indicating that the default voice might not be optimal to use in interaction.

The Google Cloud text-to-speech supported voices *en-US-Neural2-F* (female voice) and *en-US-Neural2-J* (male voice) were chosen, because of the clarity of the articulation and minimal background sound and its neutral usage purpose (e.g. not limited to news reporting). Both voices were identical in intonation, accentuation and rhythm. The Google Cloud text-to-speech API was used to generate text to synthetic voice audio files, which could then be pronounced by the Pepper robot. However, to ensure clarity, the voice audio was synced with the videos afterwards, during the editing phase.

4.3.2 Pretest 2: Speech act manipulation

Formulation of initial scenarios: The Pepper robot's speech act is manipulated by changing the formulation of the spoken scenario. Two scenario's are formulated: (1) a scenario in which the robot uses assertive speech, and (2) a scenario in which the robot uses affiliative speech. Both scenarios are created by following the distinction of Leaper & Ayres (2007) to obtain textual elements indicating assertive speech (task-oriented statements, e.g. informative suggestions, directive language, disagreeing with or criticizing the other's contributions) and affiliative speech (collaborative statements, e.g. supportive language, expressing agreement, acknowledging the other's contributions).

To ensure that these created scenarios are perceived as either assertive speech or affiliative speech, both scenarios are pretested. The goal was to determine if the affiliative scenario showed significant differences in the affiliativeness dimension (as opposed to the assertiveness dimension)

(1): Hello. Yesterday, a coworker of mine sent out a wrong e-mail, resulting in the loss of an important client of the company. My coworker explained to me what went wrong, and suggested a solution. **I told my coworker that I disagreed with the solution and that I did not think it would work. I suggested checking e-mails before sending them in the future. Then, I told my coworker that I would arrange a meeting with the client, because the problem needed to be solved right away.**

(2): Hello. Yesterday, a coworker of mine sent out a wrong e-mail, resulting in the loss of an important client of the company. My coworker explained to me what went wrong, and suggested a solution. **I told my coworker that I agreed with the solution and that I thought it was a good first proposal. I said that this could happen to anyone and that I understood how these mistakes can happen. Then, I told my coworker that we could sit down and solve the problem together.**

Figure 2: The two written scenarios in the workplace setting. (1) depicts the assertive speech act scenario and (2) the affiliative speech act scenario.

and if the assertive scenario showed significant differences in the assertiveness dimension (as opposed to the affiliativeness dimension). If not, the specific traits could be used to potentially tailor the scenarios after the pretest to align with the definitions of assertive and affiliative speech on those specific traits.

To minimize the impact of other factors such as the robot's appearance, voice and gender cue on the perception of the speech act, the scenarios are presented as written textual elements. The scenario's are identical, except for the part where the speech act is manipulated. The scenarios have equal sentence and word counts. In Figure 2, the two scenarios are shown. (1) presents the assertive speech act scenario and (2) the affiliative speech act scenario. The bold text denotes the sections of the scenarios that differ from each other.

Setting of scenarios: The chosen setting of the scenarios is a workplace. The reason behind the choice for the scenario of the robot operating in a workplace setting is threefold. First, the scenario is relatable for most participants. Second, both assertive and affiliative speech are a common communication style in this setting. Third, the relevance of my research evolves around implications of gender-emotion stereotypes, often experienced in workplace settings. This scenario is loosely based on the scenario in the experiment of Chita-Tegmark et al. (2019), which entailed three agents in an office setting. In my setup, the scenario is intentionally designed to exclude the presence of another agent, to avoid potential confounding effects related to the gender of the other agent. Thus, in the scenarios of this experiment, the speaker recounts an incident involving a coworker and how they managed the situation, while the particular coworker is not present in the scenario.

Participants: A total of 50 participants (25 female, 25 male) between 19 and 65 years old ($M = 27.73$, $SD = 8.67$) participated in the second pretest. The online research platform Prolific

was used to recruit the participants. It is important to note that the demographics information, specifically the sex of the participants was collected by Prolific using the term 'sex' (with two choices: male or female) instead of 'gender'.

The participant pool is drawn exclusively from European countries. This decision was made to ensure a relatively homogeneous sample of high quality, and because of the availability of various reports on attitude towards robots conducted by the Eurobarometer on behalf of the European Union (EU) Institutions. This makes EU countries (as well as other European countries) a suitable choice. Furthermore, all participants were required to be fluent in the English language, to ensure a comprehensive understanding of the written text. Participants were incentivized with a financial compensation for their time by Prolific.

Procedure: The participants of the second pretest were randomly assigned to one of the two scenarios. The participants were asked to read the scenario carefully and answer fourteen questions on a 7-Point Likert scale, ranging from *Strongly disagree* to *Strongly agree*. The scenario remained visible on top of the page, allowing the participants to review the text while answering the questions. The participants assessed the presented scenario across fourteen statements, in random order, related to the different components of the definitions of each speech act. The questions started with the sentence: "After reading the scenario, I believe that the speaker..", followed by the seven traits related to the assertive speech act (e.g. gives opinions) and the seven traits related to the affiliative speech act (e.g. shows support to the coworker) (as detailed in Appendix B.1).

Results: Both scenarios were evaluated on the seven traits related to assertive speech and the seven traits related to affiliative speech. Each participant evaluated one scenario, resulting in 25 participant responses per scenario. Two means were calculated per participant, combining the seven traits associated with the assertive speech act into one mean value for the assertiveness dimension, and the seven traits associated with the affiliative speech act into one mean value for the affiliativeness dimension. This creates two distributions of average participant responses per scenario. A visual representation of the findings can be found in Appendix B.2.

The first scenario (the assertive scenario) showed a high median on the assertive dimension and a low median on the affiliative dimension, while the second scenario (the affiliative scenario) showed a high median on the affiliative dimension and a low median on the assertive dimension. To determine the significance of these results, statistical tests were conducted. First, the normality and homogeneity of variances assumptions were examined using the Shapiro-Wilk test and Levene's test, respectively. The Shapiro-Wilk test indicated that one of the distributions significantly deviated from a normal distribution ($p = .013$), suggesting non-normality in the data. Therefore, the non-parametric test Mann-Whitney U for independent samples was conducted twice to conclude the second pretest results. The first test was used to assess the null hypothesis that the population medians of the two scenarios are equal on the affiliative dimension, while the second test was used to assess if they are equal on the assertive dimension. The obtained results showed significant p values for both dimensions ($p < .001$), indicating a statistically significant difference on both dimensions

between the two scenarios.

Conclusion: In conclusion, the results for the second pretest showed that the assertive scenario had significantly high ratings on the assertiveness dimension and low ratings on the affiliativeness dimension, while the affiliative scenario had significantly high ratings on the affiliativeness dimension and low ratings on the assertiveness dimension. These findings suggest that the initially written scenarios were perceived as the intended speech act. Thus, the scenarios can be used in the main experiment to manipulate the speech act of the Pepper robot.

4.4 Phase 2: Main experiment

An online, video-based experiment was designed, in which the participants watched a video of the Pepper robot. The Pepper robot is manipulated in gender by changing its voice to a female or male voice, created using the Cloud text-to-speech API, described in Pretest 1. The speech act of the Pepper robot is manipulated by changing the scenario uttered by the robot in the video, described in Pretest 2. Thus, four versions of the video are created, as a result of the two different manipulations. The participants are randomly assigned to one of the four experiment videos. In the video, the Pepper robot speaks (either with a male voice or with a female voice) about an event that occurred the day before in the office with a coworker and how it reacted to that situation (either utilizing assertive speech or affiliative speech). After watching the video, a survey containing the dependent measures is filled in by the participants.

The following sections outline the methodology of this experiment. First, the creation of the four versions of the experiment videos is described. Then, all experimental measures and questionnaire items incorporated in the survey are set out. Next, the procedure of the experiment is presented as well as the recruitment and demographics of the participants. Lastly, the process of the data preprocessing and analysis is described.

4.4.1 Experiment videos

For the main experiment, four versions of the video were created of the Pepper robot incorporating the two manipulations. The Pepper robot is shown from the waist up to prevent accidental gender attribution of the waist and hip area. The background of the videos is neutral to prevent visual distractions (as shown in Figure 1). Two videos were recorded of the Pepper robot, one in which the Pepper robot utters the assertive speech act scenario (as shown in Figure 2.1) and one video in which the Pepper robot utters the affiliative speech act scenario (as shown in Figure 2.2). In both scenarios, the sentence *"I am a social robot, I work here at the company."* is added after "*Hello.*" to ensure understanding of the context of the scenario.

The Pepper robot pronounced the words in the scenarios using the function within the SIC framework named *say_animated*. This function enables automatic hand and head gestures, that correspond with the words provided as a string argument to the function. However, *say_animated* uses the Pepper robot's default voice. As concluded from the results from the first pretest, female

and male voice files, created using the Cloud text-to-speech API, were used instead of the default voice. Therefore, during the editing phase, the sound is removed from the video, and the female and male voice files are carefully synced with the videos (matching the hand and head gestures of the Pepper robot). The clarity of the voice files is also ensured through this process.

The four different videos are created by syncing both the female and male voice files with the assertive and affiliative scenario videos. The videos are subtitled, to ensure complete understanding. All four videos are the same in all aspects, except for the manipulations. The length of the video (approximately 38 seconds), the background, the lighting, the type and the amount of gestures and the speaking rate are identical. Thus, four experiment videos were created, manipulating solely the gender (male and female) and the speech act (assertive speech and affiliative speech) of the Pepper robot.

Additionally, a video is recorded in which the Pepper robot is shown without sound. In this video, with the same setup as the experiment videos and with a length of 36 seconds, the Pepper robot performs some standard humanlike gestures (i.e. waving, bowing, looking around and lifting its arms). The goal of this video is to address the novelty effect, which refers to the tendency of people to react differently to things they experience for the first time, as advised by Hoffman & Zhao (2020) in their guidelines for HRI research. Therefore, before showing the participants the experiment video, this video is shown to mitigate the novelty effect while watching the experiment video.

The videos are uploaded and embedded using HTML code in the experiment survey. The survey is created in the secured online Qualtrics environment, which is hosted on servers authorized by Utrecht University. The survey was designed in accordance with the guidelines of Spiel et al. (2019), that indicate the best practices of surveying gender, relevant across a wide set of fields. Specifically, Spiel et al. (2019) recommend to take into account the complexity of the participants' gender identities when asking about the gender the participant identifies with. The following recommended list of five options was included: woman, man, non-binary, prefer not to disclose and prefer to self-describe (opening up a free text field) (Spiel et al., 2019). Furthermore, as recommended, the response to this question was not forced. The next section provides a detailed explanation of the measures used in the survey.

4.4.2 Experimental measures

In the experiment survey, the participants were asked to respond to several existing questionnaires. The dependent variables, indicated by the evaluation of the Pepper robot, are measured in terms of warmth, competence and discomfort. These three dimensions are combined into one validated measure, the 'Robotic Social Attributes Scale' (RoSAS). The RoSAS is developed by Carpinella et al. (2017), and is widely used to assess people's perception of the social attributes of robots. In this experiment, the RoSAS is used to determine the perception of the participants of the Pepper robot in the video in terms of warmth, competence and discomfort.

Each of three dimensions consists of six items, resulting in 18 items in total. The internal consistency of the three dimensions warmth (α = .85), competence (α = .85) and discomfort (α = .83) was measured using Cronbach's alpha. The participants were asked the following question (slightly adapted from Carpinella et al. (2017)): "Using the scale provided, how closely is the word below associated with the robot in the video?", followed by each of the 18 items. All items (listed for ease in Table 1) were presented to the participants on a 7-Point Likert Scale, ranging from *Not at all* to *Very much so*. The order of the items was randomized to prevent systemic response bias.

Warmth	**Competence**	**Discomfort**
Happy	Capable	Scary
Feeling	Responsive	Strange
Social	Interactive	Awkward
Organic	Reliable	Dangerous
Compassionate	Competent	Awful
Emotional	Knowledgeable	Aggressive

Table 1: The three dimensions of the Robotic Social Attributes Scale (RoSAS), assessing social attributes of robots as ascribed by people.

Additionally, since participants' overall acceptability and attitude towards robots, participants' view on gender roles in society and their potential ambivalent sexist attitudes were expected to impact their responses, these were considered as confounding variables. To gain insights about the background and views of the participant sample, participants' previous experience with robots and participants' opinion on robots in the workplace are assessed.

To assess participants' overall acceptability and attitude towards robots, 10 items were extracted from the General Attitudes Towards Robots Scale (GAToRS). This scale, developed by Koverola et al. (2022), measures people's attitudes towards robots on personal and societal level. The 10 societal level items are used in this experiment to assess the positive attitudes (e.g. "Robots can make life easier.") as well as negative attitudes (e.g. "Widespread use of robots is going to take away jobs from people.") of the participants about robots in general. The internal consistency was measured for both the GAToRS positive subscale (α = .75) as the GAToRS negative subscale (α = .70). All items of the GAToRS are randomized and presented on a 7-Point Likert scale, ranging from *Strongly disagree* to *Strongly agree*.

To assess participants' view on gender roles in society and their ambivalent sexist attitudes, the 12 items of the short version of the Ambivalent Sexism Inventory (ASI), developed by Glick & Fiske (1996) were used. This scale was also used to assess sexist attitudes in the study by Bernotat et al. (2017). The 12 items (α = .89), randomized and presented on a 7-Point Likert scale from *Strongly disagree* to *Strongly agree*, reflected both benevolent sexism (e.g. "Women, compared to men, tend

to have a superior moral sensibility.") as well as hostile sexism (e.g. "Women exaggerate problems they have at work.").

To assess participants' previous experience with robots, two items are used. First, "Please indicate how knowledgeable you are about robots and/or the robotics domain." (5-Point Likert scale, ranging from *Not at all knowledgeable* to *Extremely knowledgeable*). Second, "Please indicate how often you have encountered robots in the last year." (5-Point Likert scale, ranging from *Never* to *All the time*). To assess participants' opinion on robots in the workplace, one item is used, "Please indicate how you would feel about having a robot assist you at work.", presented on a 5-Point Likert scale, ranging from *Extremely negative* to *Extremely positive*. This question is adapted from the 1-item survey researching robot acceptance at work (RAW) in EU countries, as presented in the study by Turja & Oksanen (2019).

4.4.3 Procedure

The participants took part in the experiment by completing the survey online. First, in accordance with the Ethics guidelines of Utrecht University, the participants were asked explicit consent for participation and information regarding the confidentiality of their data was provided. They consented that they were 18 years or older and that they acknowledged all contents of the consent page (see Appendix C for the complete contents of the consent page). The participants who did not indicate their consent were automatically directed to the end of the survey.

The participants were provided with a brief explanation of the research goal, the survey procedure and the amount of questions to be answered. They were informed that the research goal is to examine the evaluation of a social robot functioning as a coworker in a workplace environment. Then, the participants watched the novelty effect video of the Pepper robot, which was accompanied by the explanation that they could become familiar with the robot's movements, gestures and overall capabilities. Next, the participants were notified that the next video would contain sound and instructed to either move to a quiet room or use headphones to ensure clarity of the sound. Then, the participants watched the experiment video corresponding to their experimental condition. The experiment video is introduced by the following text: "*The video below takes place in a workplace setting. The social robot is speaking about an event that occurred the day before at the office. The robot explains briefly what happened and how it reacted in that situation.*" The goal of the introduction is to ensure understanding of the context of the video.

After watching the two videos, the participants were asked to answer 19 questions about their impression of the social robot in the videos and then 27 general questions. The first question represented the attention check. This question was incorporated in the survey, to enable detection of data invalidation due to inattentiveness by the participants. The participants were asked to answer one question, designed to check whether they were fully engaged while watching the videos. This question was formulated as follows: "*About whom was the robot speaking in the video?*", including the following answer options: *"A coworker", "A friend", "A family member", "I don't*

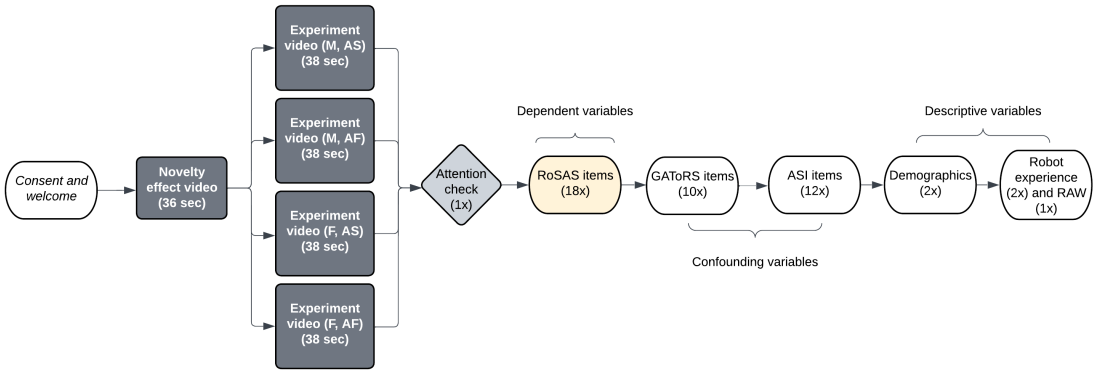


Figure 3: A flowchart explaining the time at which the experiment videos and survey questions are presented to the participant. Participants are randomly divided over the four different experiment videos with two manipulations: Robot gender (M: Male, F: Female) and Robot speech act (AS: Assertive speech, AF: Affiliative speech).

know". Participants that answered an option other than *"A coworker"* were excluded.

The other 18 of the 19 questions were the RoSAS items. Then, participants were asked to answer 10 questions "on their perception of robots in society" (the GAToRS), 12 "on their perception of gender roles in society" (the ASI), 2 on demographics (gender and age), 2 on previous experience with robots and 1 on their opinion of robots in the workplace (the RAW item). Once the survey was completed, participants were thanked and redirected to Prolific. Their data was saved and stored on the password-protected University authorized servers. The complete survey flow is shown in Figure 3.

4.4.4 Participants

Since the experiment design is created as a 2 (robot gender: male and female) x 2 (speech act: assertive speech and affiliative speech) between-subjects design, while also incorporating the participant's gender (male or female), eight participant groups are formed in order to test the four hypotheses. A power analysis was performed to obtain the desired number of participants per group. A required sample size of 171 participants was computed with a fixed effects ANOVA with eight groups, a standard error probability ($\alpha = .25$), a standard power (1 - $\beta = .9$) and estimating a medium effect size index (f = .25, $\eta^2 \approx .058$). The values used here are comparable to those presented in the study by Saunderson & Nejat (2020). The degrees of freedom (df) were calculated by the following formula, indicating a focus on the interaction effect between the different groups: $df = (n_1 - 1) \cdot (n_2 - 1) \cdot (n_3 - 1) \cdot ... \cdot (n_k - 1)$, where $n_1, n_2, n_3, ..., n_k$ denote the number of groups in each of the k factors of the design. Therefore, the degrees of freedom used in this experiment with 2x2x2 groups is $df = (2 - 1) \cdot (2 - 1) \cdot (2 - 1) = 1$. To ensure the minimum sample size is reached, an additional 15% was added to the sample size.

A total of 201 participants was recruited (98 female, 99 male, 3 non-binary and 1 prefer not to disclose) between 18 and 66 years old (M = 28.44, SD = 9.13) via the online research platform Prolific. Similar to the second pretest, the participant pool was drawn from European countries, participants were required to be fluent in the English language and they received a financial compensation from Prolific. Since the participant's gender is also treated as an experimental factor, and the group non-binary people is not sufficient in statistical power, the data has to be limited to the 98 female and 99 male participants, excluding the four participants indicating a non-binary gender and the participant that preferred not to disclose their gender.

On average, participants reported moderate knowledge about the robots or the robotic domain (M = 2.50, SD = 0.87) and reported encountering robots to a moderate extent in the last year (M = 2.43, SD = 0.87). Furthermore, participants expressed positive attitudes towards having a robot assist them at work (M = 3.53, SD = 0.91).

After excluding participants on the basis of their gender (n = 4) and excluding participants that failed the attention check (n = 3), the sample consists of 194 participants, resulting in the participant data for each experimental condition according to Table 2. Additionally, the data was examined on outliers in total survey duration time and on non-differentiating ratings in the participant responses (i.e. straight lining). No participants were excluded on this basis.

Robot gender	**Robot speech act**	**Participants**
M	AS	23F; 25M
M	AF	24F; 27M
F	AS	24F; 24M
F	AF	25F; 22M

Table 2: The final number of participants identifying as female (F) or male (M) analyzed in this study, presented per experimental manipulation: Robot gender (M: Male, F: Female) and Robot speech act (AS: Assertive speech, AF: Affiliative speech).

4.4.5 Data analysis

To test the four hypotheses stated in Section 3, a series of the Analysis of Covariance (ANCOVA) statistical tests were performed on the participant data. The dependent variables are the three dimensions warmth, competence and discomfort of the RoSAS. Replicating Carpinella et al. (2017), the six items comprising each of the three dimension were averaged for each participant. In Table 3 the means and standard deviations for each dimension are shown per condition.

For each hypothesis test, correlations with the confounding variables were tested and incorporated in the models if assumptions were met. The confounding variables that were examined were the GAToRS positive subscale (one average value of the five items), the GAToRS negative subscale

(one average value of the five items) and the ASI (one average value of the twelve items). The assumption for homogeneity of the regression slopes across the different experimental conditions is confirmed for all potential confounding variables and independent variables. Thus, there were no significant interactions between the confounding variables and the independent variables. Furthermore, correlations between the confounding variables were tested using Pearson's correlation coefficient, to assess potential relationships that could affect the accuracy and interpretation of the model's results. The analysis revealed no strong correlations among these variables. Lastly, the assumption of the existence of a linear relationship between the covariate and the dependent variable is tested for each pair of covariate and dependent variable. In Appendix D visualizations are presented of the simple linear regression models between the covariates and the dependent variables. The p values associated with these models were used to determine whether the covariates were controlled for in the model.

To test hypothesis 1, "*The female-voiced assertive Pepper robot will receive lower scores in the warmth dimension than the male-voiced assertive Pepper robot.*", a two-way ANCOVA was performed, testing the effects of the two independent variables (robot gender: male or female, robot speech act: assertive or affiliative speech) on the dependent variable warmth, while controlling for the GAToRS positive subscale. It was confirmed that the ANCOVA test assumptions were met using the Shapiro-Wilk test for normality and the Levene's test for equality of variances. Furthermore, it was confirmed that there is a linear relation between the GAToRS positive subscale and the dependent variable warmth using a linear regression model ($p = .027$). As the GAToRS negative subscale ($p = .310$) and the ASI ($p = .680$) did not meet this assumption, the model for hypothesis 1 was not controlled for these variables. The two-way ANCOVA was used to determine if there is a significant interaction effect between robot gender and robot speech act on the warmth dimension, and specifically if there is a significant difference between the female-voiced assertive robot condition and the male-voiced assertive robot condition.

To test hypothesis 2, "*The female-voiced Pepper robots will receive lower scores in the competence dimension than the male-voiced Pepper robots, independent of speech act.*", a two-way ANCOVA was performed, testing the effects of the two independent variables on the dependent variable competence, while controlling for the GAToRS positive subscale. It was confirmed that the ANCOVA test assumptions for normality and equality of variances are met. Linearity between the GAToRS positive subscale and the dependent variable competence was confirmed using a linear regression model ($p < .001$). Similar to the model of hypothesis 1, the GAToRS negative subscale ($p = .159$) and the ASI ($p = .940$) did not meet the assumption of linearity and were not controlled for in the model for hypothesis 2. The two-way ANCOVA was used to determine if there is a significant interaction effect between robot gender and robot speech act on the competence dimension. Furthermore, it was examined if there is a significant main effect of robot gender on the competence dimension, independent of robot speech act.

To test hypothesis 3, "*Participants that identify as male will give lower scores in the competence*

dimension to the female-voiced Pepper robots than participants that identify as female.", a two-way ANCOVA was performed, testing the effects of the two independent variables robot gender and participant gender on the dependent variable competence, while controlling for the GAToRS positive subscale. Since the assumptions are the same as in hypothesis 2, the GAToRS negative subscale and ASI were also not controlled for this hypothesis. The two-way ANCOVA was used to determine if there is a significant interaction effect between robot and participant gender on the competence dimension, and specifically if there is a significant difference between the ratings of participants that identify as male and participants that identify as female on the female-voiced robot condition.

To test hypothesis 4, "*The female-voiced assertive Pepper robot and the male-voiced affiliative robot will receive higher scores in the discomfort dimension, than the female-voiced affiliative Pepper robot and the male-voiced assertive Pepper robot.*", a two-way ANCOVA was performed, testing the effects of the two independent variables robot gender and robot speech act on the dependent variable discomfort, while controlling for the GAToRS positive and negative subscale. Results of the Shapiro-Wilk test ($p < .001$) and the Levene's test ($p = .034$) indicated violations of the normality equality of variances assumptions. However, it has been found that ANCOVA is relatively robust to deviations from normality (Glass et al., 1972). This is due to the central limit theorem, which states that when the sample is large enough, the means of the samples tend to follow an approximately normal distribution, even if the sample itself is not normally distributed. Therefore, ANCOVA analysis can still yield reliable results for this hypothesis.

For this model, both the GAToRS positive and the GAToRS negative subscales are taken into account, as linear relations were confirmed with the usage of linear regression models, between the dependent variable discomfort and the positive subscale ($p < .001$), as well as the the negative subscale ($p = .004$). Since no linear relationship was observed between ASI and discomfort ($p = .386$), the model did not account for ASI as a covariate. The two-way ANCOVA was used to determine if there is a significant interaction effect between robot gender and robot speech act on the discomfort dimension, and specifically the female-voiced assertive robot and the male-voiced affiliative robot were compared with the other two conditions.

	Male robot		Female robot	
	Assertive	Affiliative	Assertive	Affiliative
Dependent variables	*Means (SD)*	*Means (SD)*	*Means (SD)*	*Means (SD)*
Warmth	3.31 (1.17)	3.74 (1.21)	3.44 (0.97)	3.72 (1.18)
Competence	5.26 (0.95)	5.11 (0.88)	5.57 (0.80)	5.29 (1.06)
Discomfort	2.78 (1.35)	2.43 (0.85)	2.51 (0.90)	2.50 (0.99)

Table 3: The means and standard deviations of each condition, calculated per dimension of the RoSAS (warmth, competence, discomfort).

5 Results

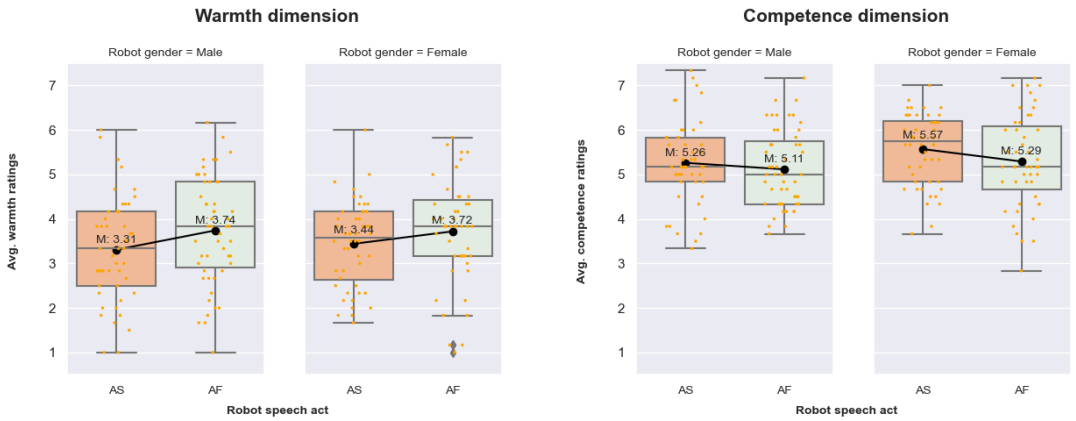
In this section, the results are presented from a series of two-way ANCOVA statistical tests conducted on the experiment data. The results are organized in four sections, each corresponding to one of the hypotheses concerning the impact of the independent variables robot gender, robot speech act and participant gender on the dependent variables warmth, competence and discomfort.

5.1 Effects of robot gender and speech act on warmth

A two-way ANCOVA was performed to examine the presence of a significant interaction effect between robot gender and robot speech act on the warmth dimension of the RoSAS, while controlling for the positive subscale of the GAToRS. The results of the two-way ANCOVA indicated that there was no significant interaction effect for robot gender and robot speech act on perceived warmth ($F(1, 189) = 0.23$, $p = .634$, $\eta^2 = .001$). Furthermore, no significant main effects were found for robot gender ($F(1, 189) = 0.14$, $p = .707$, $\eta^2 < .001$) or robot speech act ($F(1, 189) = 3.84$, $p = .052$, $\eta^2 = .019$).

The statistical results for the ANCOVA model, along with the results for the ANOVA model excluding the positive subscale of the GAToRS (GAToRS_p), are presented in Table 4. When GAToRS_p was excluded from the model, a significant main effect of robot speech act on warmth was observed ($F(1, 190) = 4.79$, $p = .030$, $\eta^2 = .025$). These results suggest that without controlling for the GAToRS_p, robot speech act influenced the perceived warmth of the robot independently from robot gender. Examining the means for the warmth dimension in Table 3, it is evident that the means for the affiliative robot conditions (male: 3.74, female: 3.72) were higher compared to the means for the assertive robot conditions (male: 3.31, female: 3.44). In Figure 4a the means are plotted for each experimental condition, together with the participants' warmth ratings.

However, when the covariate GAToRS_p was included in the ANCOVA model, the effect of robot speech act on warmth became non-significant. This suggests that participants' positive attitudes towards robots on societal level accounts for a portion of the variance in the warmth ratings. By accounting for the influence of this covariate, the true impact of the independent variables on warmth was clarified. Hypothesis 1, "*The female-voiced assertive Pepper robot will receive lower scores in the warmth dimension than the male-voiced assertive Pepper robot.*" is therefore rejected, as no significant interaction effect nor main effects were found of the independent variables robot gender and robot speech act on the participants' warmth ratings.



(a) Robot gender and speech act vs. warmth

(b) Robot gender and speech act vs. competence

Figure 4: Participant ratings in the warmth and competence dimension per experimental condition in the 2 (robot gender: male and female) x 2 (robot speech act: assertive [AS] vs affiliative [AF]) design. Means are plotted per condition.

5.2 Effects of robot gender and speech act on competence

A two-way ANCOVA was performed to investigate the presence of a significant interaction effect between robot gender and robot speech act on the competence dimension of the RoSAS, while controlling for the GAToRS_p. The results of the two-way ANCOVA indicated that there was no significant interaction effect between robot gender and robot speech act on perceived competence ($F(1, 189) = 0.27$, $p = .602$, $\eta^2 = .001$). However, simple main effect analysis showed that both robot gender ($F(1, 189) = 3.92$, $p = .049$, $\eta^2 = .002$) and robot speech act ($F(1, 189) = 4.48$, $p = .036$, $\eta^2 = .002$) did have a significant main effect on competence. These results suggest that perceived competence is affected by both robot gender and robot speech act, independent of each other, but interaction or modification of each other's effects does not occur. Specifically, participants ascribe higher competence to robots in the assertive speech act conditions compared to the affiliative speech act conditions, and participants ascribe higher competence to the female-voiced robots compared to the male-voiced robots (see Figure 4b).

The results for the ANCOVA model, along with the results for the ANOVA model excluding the GAToRS_p, are provided in Table 4. The main effects of the independent variables became significant when including GAToRS_p in the model and removing its influence on the variance in the data of the competence dimension. Hypothesis 2, "*The female-voiced Pepper robots will receive lower scores in the competence dimension than the male-voiced Pepper robots, independent of speech act.*" is rejected, as it is found that participants tend to give higher competence ratings to the female-voiced robots compared to the male-voiced robots, independent of speech act.

ANCOVA Model Hypothesis 1	η^2	sum_sq	df	F	p
C(Robot_Gender)	<.001	0.18	1	0.14	.707
C(Robot_Speech_act)	.019	4.88	1	3.84	.052
C(Robot_Gender):C(Robot_Speech_act)	.001	0.29	1	0.23	.634
GAToRS_p	.020	5.06	1	3.98	**.048**
Residual	-	240.33	189	-	-
(ANOVA Model Hypothesis 1)	η^2	sum_sq	df	F	p
C(Robot_Gender)	.006	0.14	1	0.11	.740
C(Robot_Speech_act)	.025	6.19	1	4.79	**.030**
C(Robot_Gender):C(Robot_Speech_act)	.001	0.29	1	0.22	.638
Residual	-	245.38	190	-	-

ANCOVA Model Hypothesis 2	η^2	sum_sq	df	F	p
C(Robot_Gender)	.018	3.10	1	3.92	**.049**
C(Robot_Speech_act)	.021	3.54	1	4.48	**.036**
C(Robot_Gender):C(Robot_Speech_act)	.001	0.22	1	0.27	.602
GAToRS_p	.078	13.12	1	16.61	**<.001**
Residual	-	149.25	189	-	-
(ANOVA Model Hypothesis 2)	η^2	sum_sq	df	F	p
C(Robot_Gender)	.017	2.84	1	3.32	.070
C(Robot_Speech_act)	.013	2.16	1	2.53	.114
C(Robot_Gender):C(Robot_Speech_act)	.001	0.21	1	0.25	.618
Residual	-	162.37	190	-	-

Table 4: Summary of the two-way ANCOVA results for Hypothesis 1 and 2. The two-way ANOVA models, excluding the covariate GAToRS_p, are shown below each ANCOVA model. The abbreviation C represents a categorical variable in formula notation.

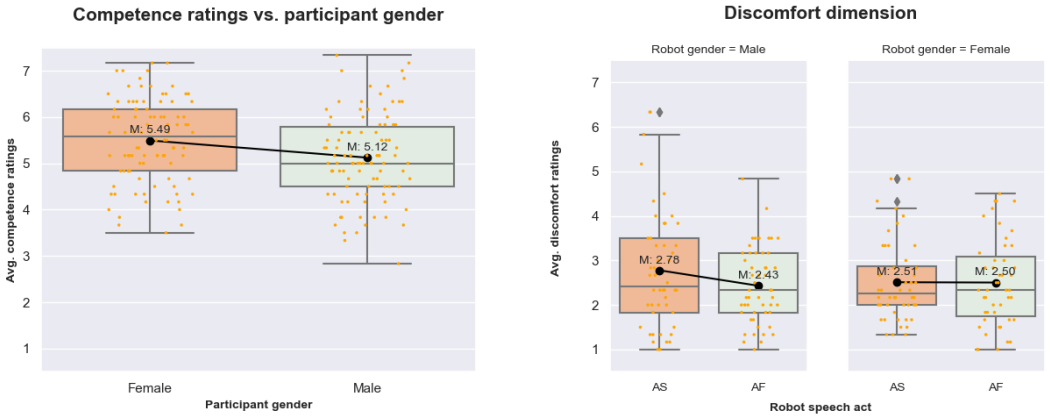
5.3 Effects of robot gender and participant gender on competence

A two-way ANCOVA was conducted to explore the presence of a significant interaction effect between robot gender and participant gender on the competence dimension of the RoSAS, while controlling for GAToRS_p. The two-way ANCOVA results indicate that there is no significant interaction effect between robot gender and participant gender on perceived competence ($F(1, 189) = 0.09$, $p = .768$, $\eta^2 < .001$). Moreover, this analysis revealed no significant main effect between robot gender and competence ($F(1, 189) = 3.71$, $p = .055$, $\eta^2 = .016$). However, a significant main effect was observed between participant gender and competence ($F(1, 189) = 14.49$, $p < .001$, $\eta^2 = .063$). This suggests that participant gender influences the perceived competence of the robot independently of the robot's gender.

The two-way ANCOVA results are presented in Table 5, along with the results for the ANOVA model excluding the GAToRS_p. The inclusion of GAToRS_p did not affect the significance of the effect of participant gender on competence. In Figure 5a, the distribution of competence ratings for participants that identify as female and participants that identify as male is visualized. It becomes evident that participants that identify as female ($M = 5.49$, $SD = 0.89$) tend to give higher ratings

than participants that identify as male ($M = 5.12$, $SD = 0.94$).

Hypothesis 3, "*Participants that identify as male will give lower scores in the competence dimension to the female-voiced Pepper robots than participants that identify as female.*" is rejected, as there was no significant interaction effect between the gender of the robot and the gender of the participant on the scores in the competence dimension. However, it was observed that the participants identifying as male tend to give lower scores in the competence dimension than participants identifying as female, but this effect is independent of robot gender.



(a) Competence ratings vs. participant gender

(b) Robot gender and speech act vs. discomfort

Figure 5: Participant ratings in the competence and discomfort dimensions. Competence ratings are shown for each participant gender (male and female), discomfort ratings are shown per experimental condition in the 2 (robot gender: male and female) x 2 (robot speech act: assertive [AS] vs affiliative [AF]) design. Means are plotted for each condition.

5.4 Effects of robot gender and speech act on discomfort

A two-way ANCOVA was performed to determine if there is a significant interaction effect between robot gender and robot speech act on the discomfort dimension, while controlling for GAToRS_p and GAToRS_n. The two-way ANCOVA results indicated that there was no significant interaction effect ($F(1, 188) = 1.82$, $p = .179$, $\eta^2 = .009$), nor significant main effects for robot gender ($F(1, 188) = 0.36$, $p = .548$, $\eta^2 = .002$) and robot speech act ($F(1, 188) = 0.32$, $p = .574$, $\eta^2 = .002$) on the discomfort dimension.

In Table 5 the ANCOVA results and the ANOVA results without the inclusion of GAToRS_p and GAToRS_n are presented. Incorporating the covariates in the model did not lead to changes in the significance of the results. In Figure 5b the participants' discomfort ratings are plotted for each experimental condition.

Hypothesis 4, "*The female-voiced assertive Pepper robot and the male-voiced affiliative robot will receive higher scores in the discomfort dimension, than the female-voiced affiliative Pepper robot and the male-voiced assertive Pepper robot.*", is rejected, since no significant interaction effect is found between robot gender and robot speech act on the participants' discomfort scores.

ANCOVA Model Hypothesis 3	η^2	**sum_sq**	**df**	**F**	***p***
C(Robot_Gender)	.016	2.79	1	3.71	.055
C(Participant_Gender)	.063	10.89	1	14.49	**<.001**
C(Robot_Gender):C(Participant_Gender)	<.001	0.066	1	0.09	.768
GAToRS_p	.095	16.35	1	21.75	**<.001**
Residual	-	142.05	189	-	-
(ANOVA Model Hypothesis 3)	η^2	**sum_sq**	**df**	**F**	***p***
C(Robot_Gender)	.016	2.60	1	3.11	.079
C(Participant_Gender)	.038	6.31	1	7.57	**.007**
C(Robot_Gender):C(Participant_Gender)	<.001	0.032	1	0.04	.846
Residual	-	158.40	190	-	-

ANCOVA Model Hypothesis 4	η^2	**sum_sq**	**df**	**F**	***p***
C(Robot_Gender)	.002	0.35	1	0.36	.548
C(Robot_Speech_act)	.002	0.31	1	0.32	.574
C(Robot_Gender):C(Robot_Speech_act)	.009	1.77	1	1.82	.179
GAToRS_p	.072	14.66	1	15.11	**<.001**
GAToRS_n	.027	5.49	1	5.65	**.018**
Residual	-	182.40	188	-	-
(ANOVA Model Hypothesis 4)	η^2	**sum_sq**	**df**	**F**	***p***
C(Robot_Gender)	.002	0.45	1	0.41	.521
C(Robot_Speech_act)	.008	1.57	1	1.45	.229
C(Robot_Gender):C(Robot_Speech_act)	.007	1.40	1	1.29	.257
Residual	-	205.36	190	-	-

Table 5: Summary of the two-way ANCOVA results for Hypothesis 3 and 4. The two-way ANOVA models, excluding the covariate GAToRS_p and GAToRS_n, are shown below each ANCOVA model. The abbreviation C represents a categorical variable in formula notation.

6 Discussion

This study expands existing HRI research regarding robot gender and the impact of norm breaking behavior on the user's evaluation by conducting an online video experiment manipulating robot gender (male and female) and robot speech act (assertive and affiliative speech). Participants evaluated the robot in terms of warmth, competence and discomfort. In this section the findings are discussed to form an answer to the research question: *"How does the gender and speech act of a social robot (Pepper) influence the evaluation of the robot in terms of warmth, competence and discomfort?"*. Furthermore, the limitations of this study are presented and suggestions for future work are given.

6.1 On warmth evaluations

Regarding people's warmth evaluations of the robot, I hypothesized that the female-voiced assertive Pepper robot would receive lower scores in the warmth dimension than the male-voiced assertive Pepper robot (H1). According to Rudman & Glick (2001) and Brescoll (2016), women engaging in agentic behavior, such as speaking assertively, contradict existing gender-emotion stereotypes, and often experience backlash in the form of social or economical penalties. Assertive men, however, behave in line with their stereotype, resulting in different evaluations for assertive men and women.

The results of the two-way ANCOVA led to the rejection of the first hypothesis, since there was no significant interaction effect, nor main effects for robot gender and robot speech act on the participants' warmth scores. Therefore, it is found that evaluation in terms of warmth is not linked to the robot's gender or speech act. Although the ANCOVA results do not support this, there seems to be a tendency in the data that people rate the affiliative robots as warmer than the assertive robots (see Figure 4a). This is not surprising, as the definition of affiliativeness (e.g. statements of support, collaboration, active understanding, agreement) is more in line with the elements of the warmth dimension (e.g. feeling, compassionate, social), than the definition of assertiveness (e.g. task-oriented statements, suggestions, critique, disagreement). The potential main effect of robot speech act on warmth evaluations was observed as significant in the ANOVA results, before the inclusion of the covariate GAToRS_p. This indicates that participants' positive attitudes towards robots have impact on warmth evaluations in such a way that the tendency of perceiving assertive robots as less warm than affiliative robots is neglected. It could imply that people with positive attitudes towards robots on societal level do not give backlash in the form of lower warmth ratings to robots speaking in an assertive way.

Nevertheless, the social or economical penalties for assertive women, as found in Psychology literature, were not automatically found in this current study's dependent variable warmth (from the RoSAS). One explanation for this finding could be that women are generally seen as more communal (i.e. warm, kind) than men (Brescoll, 2016). Thus, when encountering a female voice in the robot, the manipulated speech act could be perceived as more affiliative than when encountering a male

voice. This could mean that the assertive female robot was potentially perceived as less assertive than the assertive male robot, resulting in higher warmth ratings. Although the assertive and affiliative speech act scenarios were pretested to be perceived as such, future work could incorporate an additional manipulation check in the main experiment, to account for the additional factor of the gender of the voice uttering the scenario.

Furthermore, the backlash effect could be interpreted as a broader concept than only a penalty in warmth rating, defined here with the items *Happy, Feeling, Social, Organic, Compassionate, Emotional*. In future work, other dependent variables than warmth should be researched, to further research the backlash effect in human-robot interaction. Namely, the backlash effect could also be found in terms of salary, career success, hiring rates or willingness to collaborate with the robot in the workplace. Moreover, the scenario in this experiment, in which the robot functioned as a coworker, could be adapted to a leadership role or another role in which the robot is in charge, while carefully considering the repercussions that choice could have to a feeling of discomfort in the participant. This could imitate the social contexts as described in the research by Rudman & Glick (2001) and Brescoll (2016) more closely. It also sounds promising to explore scenarios where the robot speaks to the coworker directly, instead of the coworker being the object of conversation. Firstly, to research the interaction context in which the participant is the observer instead of interactant, as researched by Chita-Tegmark et al. (2019). And secondly, to explore the additional impact of the gender of the coworker on the robot's evaluation, as researched by Galatolo et al. (2022).

6.2 On competence evaluations

Regarding people's competence evaluations of the robot, two hypotheses were formulated. Firstly, I hypothesized that the female-voiced Pepper robot would receive lower scores in the competence dimension than the male-voiced Pepper robot, independent of speech act (H2). This was based on the research by Ragins & Winkel (2011), stating that stereotypical feminine traits do not align with the traits associated with successful leaders, such as competence. Because of this mismatch, women tend to be perceived as less competent than men. Secondly, I hypothesized that participants that identify as male would give lower scores in the competence dimension to the female-voiced Pepper robot than participants that identify as female (H3), since Hentschel et al. (2019) found that men tend to describe women as less competent than women describe women. In the next paragraphs, I will discuss the findings of these two hypotheses.

The results of the two-way ANCOVA for the second hypothesis indicated that there was no significant interaction effect between robot gender and robot speech act on the participants' competence scores, but significant main effects have been found for both robot gender and robot speech act. Namely, female-voiced Pepper robots were ascribed higher competence than male-voiced Pepper robots and assertive Pepper robots were ascribed higher competence than affiliative Pepper robots. The first finding contradicts the second hypothesis, expecting that female robots would be

perceived as less competent than male robots. This indicates that the gender-emotion stereotype of women being perceived as less competent than men is not translated to HRI in this current experiment.

Despite the fact that an opposite finding was expected based on Psychology literature, this result replicates a key finding in the study by Carpinella et al. (2017) in which they developed and validated the RoSAS. They found that female and androgynous robots were ascribed higher competence ratings (and higher warmth ratings) than the male robots, independent of the humanlikeness of the robot. A possible explanation for this finding is the fact that gendered robots already in use are predominantly designed as female (Alesich & Rigby, 2017; West et al., 2019). This could potentially bias people towards linking the usage of robots in everyday life with femininity.

The finding might also be related to the choice of scenario, in which the robot offers to help a coworker with a mistake. Potentially, helping others can be perceived as a serving role, and thus stereotypically feminine (Carli, 2001). Researchers such as Tay et al. (2014) have found that people tend to favor robots performing a task in line with existing gender stereotypes associated with that task. This could support the finding that female robots are perceived as more competent than male robots in this specific scenario. Future work should explore various scenarios to create a broader understanding on the impact of gender on competence ratings. Additionally, the items of the RoSAS competence dimension could be replaced by other competence metrics. Namely, Hentschel et al. (2019) found that there were differences in leadership competence ratings between men and women, while ratings on instrumental competence (focused on performance) did not differ. This indicates that gender can have a different impact on evaluations depending on the specific definition of competence.

The latter significant main effect of the second hypothesis, indicating that assertive robots are perceived as more competent than affiliative robots, independent of robot gender, is a less unexpected finding. First and foremost, intuitively, it could be assumed that assertiveness shows more competence than affiliativeness, as agentic behavior is often associated with leadership competence (Hentschel et al., 2019). Furthermore, this finding is supported by prior HRI research. Namely, in the research by Neuteboom & de Graaf (2021) it is found that robots were perceived as more competent when performing an analytical task (instead of a social task), independent of robot gender. In their experiment, robot gender and task type were manipulated, where the social task was more associated with femininity and the analytical task with masculinity. They found that the effect of the robot's gender was eliminated by the predominant effect of task type on the competence evaluations. Another study by Kuchenbrandt et al. (2014) found that people are less willing to accept help from a robot performing a social task than an analytical task, also independent of robot or participant gender.

So, it appears that robots performing more analytical tasks, or in the case of this current experiment, behaving more agentically by speaking in an assertive way, are more positively evaluated in terms of competence, while eliminating the possible effects of the gendered embodiment of the

robot. This relates to the fact that people in general perceive robots in a more utilitarian manner, indicating a preference of using robots for more practical reasons (Neuteboom & de Graaf, 2021).

Thus, the presence of a significant main effect of robot speech act on competence is grounded in literature, however, future research is needed on this topic as HRI researchers are striving to incorporate robotics in social contexts (also involving more social tasks or behavior). It is important that the competence evaluations of these robots remain high, despite the fact that the robots show more social behavior.

It is important to note that these findings only became significant after controlling for the GAToRS_p. In this process, the means of the different experimental groups are adjusted based on the influence of the GAToRS_p, allowing the comparison of group means to be more accurate by focusing on differences specifically related to the independent variables. The main effects of the independent variables were isolated by accounting for the influence of the GAToRS_p scores, which were positively linearly related to participants' competence ratings (see Appendix D). This indicates that participants with higher positive attitudes towards robots on societal level generally perceive robots as more competent. Thus, when accounting for this influence in the model, the participants' tendency of perceiving assertive robots as more competent than affiliative robots, and female robots as more competent than male robots, became significant.

The results for the two-way ANCOVA performed on the third hypothesis, stating that participants identifying as male would give lower competence ratings to the female robot than participants identifying as female, indicated no significant interaction effect between robot gender and participant gender on competence. This led to the rejection of the third hypothesis. However, a significant main effect of participant gender on perceived competence of the robot was found, showing the tendency of participants that identify as female to ascribe higher competence ratings to robots (independent of robot gender) than participants identifying as male.

This finding is supported by prior HRI research, such as the research by Winkle et al. (2022). They found that participant gender impacted perceived robot credibility (in terms of expertise and trustworthiness among other items) as well as perceived robot effectiveness (in terms of performing the task of enthusiast new robotics students). Namely, women found the robot more credible and more effective than men. Furthermore, in the study by Chita-Tegmark et al. (2019), it was found that participants identifying as male rated human agents higher in emotional intelligence than robot agents, while participants identifying as female rated robot agents higher than human agents. These findings imply that women tend to ascribe more trust and capabilities to robots, or potentially technology in general, than men.

Thus, it has been shown that the finding that participants identifying as female tend to give higher competence ratings replicates previous HRI research. It is worth noting that contradicting results have also been found. In the study by Kuchenbrandt et al. (2014), for example, it was found that participants identifying as female were less compliant with the robot than men, rather making their own choices instead of following the robot's instructions. They argue that this could be based

on reduced trust in the robot, directly contradicting the findings of Chita-Tegmark et al. (2019). This emphasizes that future research should further investigate the impact of participant gender on different forms of robot's evaluations, and the interplay between participant and robot gender, creating a broader view on robot perception by different users. Moreover, extension to non-binary gender identities is crucial, as I will further explore in the limitations section of this discussion.

6.3 On discomfort evaluations

Regarding people's discomfort evaluations of the robot, I hypothesized that the Pepper robots engaging in behavior mismatching gender-emotion stereotypes (the female-voiced assertive robot and the male-voiced affiliative robot) would receive higher discomfort ratings (H4). This was based on the research by Neuteboom & de Graaf (2021) and Carpinella et al. (2017), stating that when a mismatch between robot gender and associated stereotypical behavior occurs, this could lead to dehumanization and thus a greater feeling of discomfort.

The results of the two-way ANCOVA led to the rejection of the fourth hypothesis, since there was no significant interaction effect, nor main effects for robot gender and robot speech act on the participants' discomfort scores. Unlike findings of prior HRI research such as Tay et al. (2014), indicating people's preference for robots behaving corresponding to their associated gender stereotype, the findings of the current research seem to suggest that (the interplay of) robot gender and robot speech act do not directly impact the participant's feeling of discomfort. When observing the overall means of the discomfort ratings, all conditions are rated below the average. An idealistic view on this observation is that people would be positive to interacting with robots in the workplace. This corresponds to the average value of the RAW item ($M = 3.53$, $SD = 0.91$), indicating that participants expressed predominantly positive attitudes towards having a robot assist them at work.

Although not supported by the ANCOVA results, there seems to be a tendency when observing Figure 5b that participants give higher discomfort ratings to the assertive male robot condition than to the other three conditions. Also, there seems to be a difference between the mean discomfort of the assertive and affiliative male condition, while the means of the assertive and affiliative female conditions seem to be equal. So, it appears that the discomfort ratings, represented by the items *Scary, Strange, Awkward, Dangerous, Awful* and *Aggressive*, are not influenced by the speech act of the woman, while this factor may have an influence on the discomfort ratings for men.

According to the study by Carli (2001), men are expected to show higher levels of dominant behavior such as competitiveness and aggressiveness than women, and thus people are more tolerant of this behavior in men than in women. This could potentially explain why the data of the current experiment seems to suggest that the discomfort items such as *Aggressive* and *Dangerous* were rated higher for the assertive male condition, as men behaving assertively is in line with the stereotype described by Carli (2001). It could be that the female robot, whether speaking assertively or affiliatively, was less recognized as dominant or aggressive, as this does not fit the stereotype.

Furthermore, if significant, these findings would replicate the research by Carpinella et al. (2017), as they found that male robots were rated higher on discomfort than female robots. In their research, the male machinelike robot (instead of the humanlike robot) was rated highest on discomfort. One could potentially see a correlation between the level of humanlikeness and the level of affiliativeness in this experiment, supporting the finding that the male assertive robot was rated highest on discomfort, similar to the male machinelike robot. Given the lack of significant statistical evidence in the current experiment, future research should further investigate the discrepancy between the female and male conditions in terms of discomfort, in order to make statements regarding these findings.

6.4 Limitations and future directions

The current study is subject to several limitations regarding data collection, the manipulations and experimental setup. The limitations regarding data collection are threefold. First, the participant data was collected exclusively from European countries via the online research platform Prolific. Therefore, the findings of this experiment should not be universally generalized. Since gender and associated norms and expectations are not fixed, a diversity in culture, age and gender identity amongst others is of the utmost importance. Furthermore, it has been found that Europeans, in comparison to people from Asia or the United States, have more sceptical beliefs about a robot's potential cognitive and emotional capacities (de Graaf et al., 2021). Thus, replicating this type of research on a more diverse participant pool is crucial. Here, I look at studies such as Winkle et al. (2022), where they researched their prior findings in a cross-cultural replication.

Second, the average age of the participant sample was 28.44 years old. Specifically in the context of gender and HRI, the relatively young age of the sample may have had a significant impact on the obtained results, as younger generations tend to be less conservative in their understanding of gender issues and traditional gender roles (Perugia & Lisy, 2022). Also, the acceptance of robots and technology have evolved over time, further emphasizing the importance of examining different age groups.

Third, in this study I had to exclude four participants that did not identify as the binary gender identities male or female. Due to the limited scope of this study, there was insufficient data of participants identifying beyond the binary to support this analysis. I would like to emphasize the fact that future research should have greater focus on recruiting participants identifying with gender identities beyond cisgender (including non-binary, transgender, gender fluid and gender non-conforming people). It is crucial to broaden the scope of gender identities studied, especially since the third hypothesis brought to light the complex interplay between participant and robot gender.

Then, there were several limitations concerning the decisions in regards to the manipulations and experimental setup. First, the findings of this study are limited to the Pepper robot and cannot be generalized to other robot platforms. The research of Perugia et al. (2022) provided evidence for the presence of biases and varying user perceptions associated with different robotic attributes.

Since the specific characteristics of the Pepper robot have significant impact on the outcome of this study, the findings are not necessarily the same for other robotic platforms.

Second, there has been prior HRI work, such as the research by Jackson et al. (2020), stating that Pepper's morphology is implicitly feminine, which could evoke the gendering process by the user, impacting the researcher's gender manipulation. In this study, the perceived genderedness of the Pepper robot was not pretested. Although the Pepper robot was shown in the experiment videos only from the waist up, this could have influenced the gender manipulation in the experiment. Furthermore, I want to note that I did not pretest different gender cues, but rather focused on voice as a gender cue, following the guidelines by Perugia & Lisy (2022). However, it could be fruitful to pretest the usage of different or multiple gender cues in future experiments. In this experiment, the voices used as gender cue were chosen from the humanlike Google Cloud text-to-speech supported voices, created to be perceived as either male or female.

This brings me to my third point. In this study, I had to limit myself to the usage of humanlike voices, instead of the default voice provided by the Pepper robot system. This decision was based on the results of Pretest 1, which indicated that the default voice of the Pepper system could not be manipulated to an unambiguously perceived male voice. Given the importance of accurate gender manipulation, I prioritized this aspect over the usage of the Pepper system as a whole (the physique as well as the voice). The choice of humanlike voices instead of robotic voices may have influenced the obtained results. It would be worth researching what the implicit genderedness of the default voice entails for the usage of the Pepper robot in experiments. Moreover, the impact of employing humanlike voices instead of robotic voices in robots would be worthwhile to explore.

Furthermore, the experiment was performed online with video recordings of the Pepper robot. Although this has its clear benefits for the researcher (e.g. repeated viewings, identical interactions for all participants, easier access to a large participant sample), the participants were not physically interacting with the robot. It could be the case that the participants are more invested in the experiment in a lab setting. Also, they have the ability to capture more subtle details of the robot, including the exact size, sounds and specific abilities such as making eye contact. This experiment, as well as other HRI interaction experiments I came across, could benefit from replication with in-lab studies.

Lastly, it is important to note that this study was limited by the fact that it only includes a male and female gender manipulation, excluding a genderless condition. The design of genderless or gender-neutral robots has been researched within the HRI community, but the challenge remains that the gendering process is strongly intertwined with the human form (Perugia & Lisy, 2022). This challenge indicates that there is still a lot unknown about the gendering process and its implications. To create more inclusive robotic design, future work should continue researching gender within HRI to fill this knowledge gap.

7 Conclusion

In the HRI community, there has been increased awareness that humanlike design, while often enhancing interaction and acceptance, could reproduce societal bias. It has been found that equipping robots with humanlike characteristics invites social categorization and influences the user's perception of the robot. However, there is still little knowledge about the gendering process of social robots and its implications. This study attempted to address the existing gap in knowledge by conducting a behavioral experiment regarding gender-emotion stereotypes.

This study has contributed to the HRI field by offering new insights and replicating previous findings on the impact of gender, voice and emotion on the user's evaluation in terms of warmth, competence and discomfort, incorporating both robot and participant gender. The robot's gender (male or female) and the robot's speech act (assertive or affiliative speech) were manipulated in an online video setting with a 2x2 between-subjects design. In this section the findings are concluded for each of the dimensions warmth, competence and discomfort to answer the research question, *"How does the gender and speech act of a social robot (Pepper) influence the evaluation of the robot in terms of warmth, competence and discomfort?"*.

Results have shown that the evaluation in terms of warmth was not influenced by the robot's gender or speech act, although a non-significant trend in the data suggests that affiliative robots were perceived as warmer than assertive robots. The backlash effect as found in Psychology literature, indicating that women receive social or economical penalties when behaving agentically, was not reflected by the findings of this HRI experiment. Namely, the findings did not support such an effect when a female-voiced robot exhibited agentic behavior by speaking assertively. However, it is worthwhile to explore other scenarios or expand the concept of the backlash effect beyond the warmth dimension, to further examine this gender-emotion stereotype in HRI.

Significant results have been found for the evaluation in terms of competence. Female robots were ascribed higher competence than male robots, contradicting gender-emotion stereotypes as found in Psychology literature. Results also indicated that assertive robots were ascribed higher competence than affiliative robots. This replicates prior HRI studies, stating that people perceive robots performing analytical tasks as more competent than social tasks, while neglecting a potential effect of the gendered embodiment of the robot. Furthermore, it was found that participants identifying as female tend to perceive the robot as more competent than participants identifying as male.

This study also investigated the impact of robot gender and speech act on the participant's feeling of discomfort. Results indicated that there were no significant effects on the perceived discomfort. There seems to be a tendency, although not supported by significant evidence, that the assertive male robot received higher discomfort ratings than the three other conditions. Interestingly, this discrepancy in discomfort was not observed for the assertive female robot, which was equal in discomfort ratings to the affiliative female robot.

Overall, this study has shown that the robot's gender and speech act as well as the gender

of the user interacting with the robot may have a significant impact on evaluations. Specifically, this study sought to explore behavior that challenges gender-emotion stereotypes, and how this affects the perception of the interactant. Since gendered technological design becomes increasingly integrated into our daily lives, HRI researchers and designers of social robotics face a challenging but crucial task. This study aimed to emphasize the critical importance to design social robotics without the risk of reinforcing gender bias in society. The findings suggest that gender and its associated norms and expectations may have complex effects on HRI, mainly in the competence dimension. It is essential to conduct more theory-driven experiments to approach and bring to light gender issues within HRI, while taking the complexity and diversity of the gender research space into account.

Acknowledgments

First and foremost, I would like to express my gratitude towards my supervisor, Dr. Maartje de Graaf, who has provided me with great guidance and feedback throughout my entire thesis. She brought me together with Geertje Hendriks, Chandni Bagchi, Stan de Reuver and Benedetta Ghedi for our weekly 'robot lab meetings'. These were as much informative as they were fun, and I am forever thankful to my fellow students for making these eight months an overall wonderful experience. Furthermore, I would like to thank Dr. Giulia Perugia, Dr. Ruth van Veelen, Dr. Patrícia Alves-Oliviera, Dr. Hee Rin Lee and Dr. Christine Bauer for their valuable feedback in my design phase. Also, many thanks to Dr. Eelco Herder for taking the time to read and assess this thesis. Finally, I am thankful to my family and friends for their support and enthusiasm, especially my parents and siblings, of course Frank and Anouk, my fellow thesis companions Marijn and Ellen and lastly Jikkie for her amazing video editing skills.

References

Abele, A. (2003). The Dynamics of Masculine-Agentic and Feminine-Communal Traits: Findings from a Prospective Study. *Journal of personality and social psychology*, 85(4), 768.

Alesich, S. & Rigby, M. (2017). Gendered Robots: Implications for Our Humanoid Future. *IEEE Technology and Society Magazine*, 36, 50–59.

Bernotat, J., Eyssel, F., & Sachse, J. (2017). Shape It – The Influence of Robot Body Shape on Gender Perception in Robots. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9* (pp. 75–84).

Brescoll, V. L. (2016). Leading with their hearts? How gender stereotypes of emotion lead to biased evaluations of female leaders. *The Leadership Quarterly*, 27(3), 415–428.

Bryant, D., Borenstein, J., & Howard, A. (2020). 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). Cambridge United Kingdom: ACM.

Carli, L. L. (2001). Gender and Social Influence. *Journal of Social Issues*, 57(4), 725–741.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The Robotic Social Attributes Scale (RoSAS): Development and Validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 254–262). Vienna Austria: ACM.

Chita-Tegmark, M., Lohani, M., & Scheutz, M. (2019). Gender Effects in Perceptions of Robots and Humans with Varying Emotional Intelligence | IEEE Conference Publication | IEEE Xplore. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 230–238).: IEEE.

Derks, B., Van Laar, C., & Ellemers, N. (2016). The queen bee phenomenon: Why women leaders distance themselves from junior women. *The Leadership Quarterly*, 27(3), 456–469.

Eagly, A. H. & Karau, S. J. (2002). Role Congruity Theory of Prejudice Toward Female Leaders. *Psychological Review*, 109, 573–598.

Eyssel, F. & Hegel, F. (2012). (S)he's Got the Look: Gender Stereotyping of Robots. *Journal of Applied Social Psychology*, 42(9), 2213–2230.

Fink, J. (2012). Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction. In *Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012. Proceedings 4* (pp. 199–208).

Galatolo, A., Melsión, G. I., Leite, I., & Winkle, K. (2022). The Right (Wo)Man for the Job? Exploring the Role of Gender when Challenging Gender Stereotypes with a Social Robot. *International Journal of Social Robotics*, (pp. 1–15).

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of educational research*, 42(3), 237–288.

Glick, P. & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism. *J. Pers. Soc. Psychol.*, 70(3), 491–512.

Google Cloud Text-to-Speech API: Creative Commons 4.0 Attribution License, `https://cloud.google.com/text-to-speech/docs`.

de Graaf, M. M., Allouch, S., & Van Dijk, J. A. (2015). What Makes Robots Social?: A User's Perspective on Characteristics for Social Human-Robot Interaction. In *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7* (pp. 184–193).

de Graaf, M. M., Hindriks, F. A., & Hindriks, K. V. (2021). Who Wants to Grant Robots Rights? In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 38–46).

Harper, M. & Schoeman, W. J. (2003). Influences of gender as a basic-level category in person perception on the gender belief system. *Sex roles*, 49, 517–526.

Hentschel, T., Heilman, M. E., & Peus, C. V. (2019). The Multiple Dimensions of Gender Stereotypes: A Current Look at Men's and Women's Characterizations of Others and Themselves. *Frontiers in Psychology*, 10.

Hoffman, G. & Zhao, X. (2020). A Primer for Conducting Experiments in Human&Robot Interaction. *ACM Transactions on Human-Robot Interaction*, 10(1), 6:1–6:31.

Hover, Q. R. M., Velner, E., Beelen, T., Boon, M., & Truong, K. P. (2021). Uncanny, Sexy, and Threatening Robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 (pp. 119–128). New York, NY, USA: Association for Computing Machinery.

Jackson, R. B., Williams, T., & Smith, N. (2020). Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 559–567).: Association for Computing Machinery.

Koeman, V., Ligthart, M. E. U., & Hindriks, K. (2022). The Social Interaction Cloud (SIC).

Koverola, M., Kunnari, A., Sundvall, J., & Laakasuo, M. (2022). General Attitudes Towards Robots Scale (GAToRS): A New Instrument for Social Surveys. *International Journal of Social Robotics*, 14(7), 1559–1581.

Kuchenbrandt, D., Häring, M., Eichberg, J., Eyssel, F., & André, E. (2014). Keep an Eye on the Task! How Gender Typicality of Tasks Influence Human–Robot Interactions. *International Journal of Social Robotics*, 6, 417–427.

Leaper, C. & Ayres, M. M. (2007). A Meta-Analytic Review of Gender Variations in Adults' Language Use: Talkativeness, Affiliative Speech, and Assertive Speech. *Personality and Social Psychology Review*, 11(4), 328–363.

Lorber, J. (2018). The Social Construction of Gender. In *The Inequality Reader: Contemporary and Foundational Readings in Race, Class, and Gender* (pp. 347–352).

Marchetti-Bowick, M. (2009). Is Your Roomba Male or Female? The Role of Gender Stereotypes and Cultural Norms in Robot Design. *Intersect: The Stanford Journal of Science, Technology, and Society*, 2(1).

McGinn, C. & Torre, I. (2019). Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '19 (pp. 211–221). Daegu, Republic of Korea: IEEE Press.

Merkel, D. (2014). Docker: Lightweight Linux containers for consistent development and deployment. *Linux journal*, 2014(239), 2.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, 19(2), 98–100.

Moss-Racusin, C. A., Phelan, J. E., & Rudman, L. A. (2010). When Men Break the Gender rules: Status Incongruity and Backlash Against Modest Men. *Psychology of Men & Masculinity*, 11(2), 140–151.

Neuteboom, S. Y. & de Graaf, M. M. (2021). People's Perceptions of Gendered Robots Performing Gender Stereotypical Tasks. In *Social Robotics: 13th International Conference, ICSR 2021, Singapore, Proceedings 13* (pp. 24–35).

Otterbacher, J. & Talias, M. (2017). S/he's too Warm/Agentic!: The Influence of Gender on Uncanny Reactions to Robots. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 214–223). Vienna Austria: ACM.

Pandey, A. K. & Gelin, R. (2018). A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind. *IEEE Robotics & Automation Magazine*, 25(3), 40–48.

Park, G., Yaden, D. B., Schwartz, H. A., Kern, M. L., Eichstaedt, J. C., Kosinski, M., Stillwell, D., Ungar, L. H., & Seligman, M. E. P. (2016). Women are Warmer but No Less Assertive than Men: Gender and Language on Facebook. *PLOS ONE*, 11(5).

Perugia, G., Guidi, S., Bicchi, M., & Parlangeli, O. (2022). The Shape of Our Bias: Perceived Age and Gender in the Humanoid Robots of the ABOT Database. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '22 Sapporo, Hokkaido, Japan: IEEE Press.

Perugia, G. & Lisy, D. (2022). Robot's Gendering Trouble: A Scoping Review of Gendering Humanoid Robots and its Effects on HRI. Article in preparation.

Perugia, G., Rossi, A., & Rossi, S. (2021). Gender Revealed: Evaluating the Genderedness of Furhat's Predefined Faces. In *Social Robotics: 13th International Conference, ICSR 2021, Singapore, Singapore, November 10–13, 2021, Proceedings 13* (pp. 36–47).

Ragins, B. R. & Winkel, D. E. (2011). Gender, emotion and power in work relationships. *Human Resource Management Review*, 21(4), 377–393.

Reich-Stiebert, N. & Eyssel, F. (2017). (Ir)relevance of Gender? On the Influence of Gender Stereotypes on Learning with a Robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 166–176).: Association for Computing Machinery.

Richards, C., Bouman, W. P., Seal, L., Barker, M. J., Nieder, T. O., & T'Sjoen, G. (2016). Non-binary or genderqueer genders. *International Review of Psychiatry*, 28(1), 95–102.

Rudman, L. A. & Glick, P. (2001). Prescriptive Gender Stereotypes and Backlash Toward Agentic Women. *Journal of Social Issues*, 57(4), 743–762.

Saunderson, S. & Nejat, G. (2020). Investigating Strategies for Robot Persuasion in Social Human–Robot Interaction. *IEEE Transactions on Cybernetics*, 52(1), 641–653.

Siegel, M., Breazeal, C., & Norton, M. I. (2009). Persuasive Robotics: The influence of robot gender on human behavior. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2563–2568). ISSN: 2153-0866.

Soedirgo, J. & Glas, A. (2020). Toward Active Reflexivity: Positionality and Practice in the Production of Knowledge. *PS: Political Science & Politics*, 53(3), 527–531.

SoftBank Robotics America Inc. (2023). Meet Pepper: The Robot built for people. `https://us.softbankrobotics.com/pepper`.

Spiel, K., Haimson, O. L., & Lottridge, D. (2019). How to Do Better with Gender on Surveys: A Guide for HCI Researchers. *Interactions*, 26(4), 62–65.

Søraa, R. A. (2017). Mechanical genders: How do humans gender robots? *Gender, Technology and Development*, 21(1-2), 99–115.

Tajfel, H. (1979). Individuals and groups in social psychology. *British Journal of Social & Clinical Psychology*, 18(2), 183–190.

Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior*, 38, 75–84.

Taylor, S. E. (1981). A Categorization Approach to Stereotyping. *Cognitive Processes in Stereotyping and Intergroup Behavior*, (pp. 83–115).

Turja, T. & Oksanen, A. (2019). Robot Acceptance at Work: A Multilevel Analysis Based on 27 eu Countries. *International Journal of Social Robotics*, 11(4), 679–689.

West, M., Kraut, R., & Ei Chew, H. (2019). I'd blush if I could: Closing gender divides in digital skills through education. UNESCO: EQUALS Skills Coalition.

Williams, J. C., Phillips, K. W., & Hall, E. V. (2016). Tools for Change: Boosting the Retention of Women in the STEM Pipeline. *Journal of Research in Gender Studies*, 6(1), 11–75.

Winkle, K., Jackson, R. B., Melsión, G. I., Brščić, D., Leite, I., & Williams, T. (2022). Norm-Breaking Responses to Sexist Abuse: A Cross-Cultural Human Robot Interaction Study. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 120–129).: IEEE.

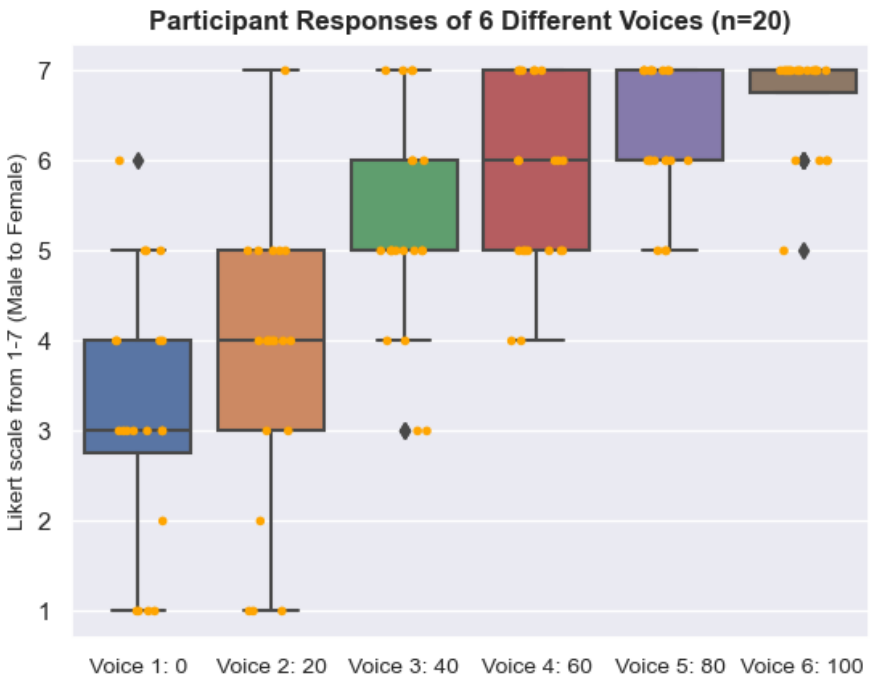
Winkle, K., McMillan, D., Arnelid, M., Balaam, M., Harrison, K., Johnson, E., & Leite, I. (2023). Feminist Human-Robot Interaction: Disentangling Power, Principles and Practice for Better, More Ethical HRI. Article in preparation.

Winkle, K., Melsión, G. I., McMillan, D., & Leite, I. (2021). Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 29–37).: Association for Computing Machinery.

Appendices

A Figures Pretest 1

The figure below shows the Pretest 1 participant responses ($n = 20$) for the six different voices, created by recording the default voice of the Pepper robot with six different pitch settings (0 to 100). Participants were asked to evaluate the voices on a 7-point Likert scale from *Male* to *Female*, with an explicitly stated neutral option.



B Figures Pretest 2

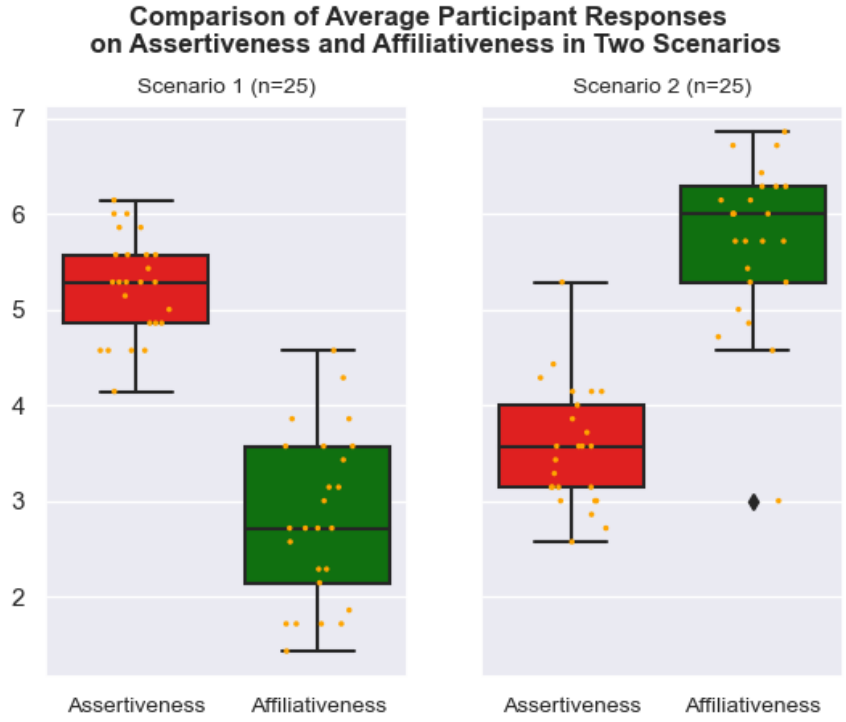
B.1 Items Pretest 2

The table below presents the fourteen statements related to the different components of the definitions of assertive speech and affiliative speech. The components were preceded in Pretest 2 by the sentence: "After reading the scenario, I believe that the speaker..".

Assertive speech act	**Affiliative speech act**
... gives out directive or task-oriented statements	... shows support to the coworker
... gives suggestions	... praises the coworker
... gives opinions	... initiates collaboration with the coworker
... gives out imperative statements (i.e. orders)	... actively tries to understand the coworker
... disagrees with the coworker	... agrees with the coworker
... promotes its own personal agency	... positively engages with the coworker
... criticizes or disapproves the contribution of the coworker	... acknowledges the contribution of the coworker

B.2 Participant responses Pretest 2

The figure below shows the 25 participant responses per scenario tested in Pretest 2. Two means were calculated per participant, combining the seven traits, associated with the assertive speech act into one mean value for the assertiveness dimension, and the seven traits associated with the affiliative speech act into one mean value for the affiliativeness dimension.



C Consent page

Welcome to the research study!

You are invited to participate in a research project conducted by Utrecht University (The Netherlands). This study is part of a master thesis project and contributes to the field of Human-Robot Interaction. If you decide to participate, you will be asked to carefully watch two short videos and answer several (self-descriptive) questions. This survey should take you around **10 minutes** to complete.

There are no anticipated risks in this study, and you should not expect to receive any direct benefits. To maintain **confidentiality**, we will assign all your data from this survey a numerical code. Your data record will not be connected to your identity. Your participation in this research is voluntary. You have the right to withdraw at any point during the study, for any reason, and without any prejudice. If you withdraw, any personal data already collected from you will be erased.

If you have questions, or if you would like to contact the researcher of this study to discuss this research, please send an e-mail to *a.i.kapteijns@students.uu.nl*. If you have any questions about your rights as a participant, please contact our Academic Integrity Counsellor by sending an email to *vertrouwenspersoon-wi@uu.nl*. If you wish, you may print a copy of this consent for your records.

By clicking the "consent"-button below, you acknowledge that the material you contribute is used to generate insights for this research project or other future publications, that your participation in the study is **voluntary**, you are **at least 18 years** of age, and that you are aware that you may choose to terminate your participation in the study at any time and for any reason. Please note that this survey can only be displayed on a laptop or desktop computer. Some features are not compatible for use on a mobile device.

Yes, I consent.

No, I do NOT consent.

D Correlations between covariates and dependent variables