RWTH AACHEN UNIVERSITY – Faculty of Arts and Humanities
Chair Individual and Technology

Robots as Allies Against Sexism in Human-Robot Groups: Can robots countering sexist comments by male-read group members emotionally support and empower female-read group members?

Master's Thesis

by

Sarah Gosten

425094

as part of the Master's Program *Computational Social Systems*

in Winter Term 2023/24

1st Examiner: Prof. Dr. Astrid Rosenthal-von der Pütten
2nd Examiner: Do Heqiu Song

Contents

List of Figures

List of Tables

Abstract

In a society where gender biases persist, technological advancements are not immune to perpetuating existing prejudices against women. While initial research was able to show the potential of social robots to counter gender stereotypes, this study seeks to investigate their role in addressing sexist remarks within group settings. Using a mixed methods laboratory approach, the study examined how people react when the social robot Pepper intervenes in sexist encounters. Participants (N = 68) engaged in a game scenario where a sexist comment was uttered, prompting Pepper to intervene in one of three ways: 1. avoidant, 2. argumentative, or 3. morally judgmental. Participants were assigned roles as either the target of the comment or the bystander. Results revealed that exposure to sexist remarks elicited negative emotions among participants. Participants who were the target of the sexist comment rated the sexist confederate significantly worse than both the robot and the bystander. This demonstrates that actively engaging in group conflicts and intervening may enhance individuals' perceptions of Pepper's suitability as a team member, potentially reaching human ratings.

Furthermore, the study found that an avoidant response from Pepper may not effectively address the sexist nature of the comment, while a confrontational approach showed promise. There are tendencies that the morally judgmental response risks escalating conflicts. Further research is necessary to verify these findings.

In conclusion, this study demonstrates the potential of social robots to intervene in sexist encounters and prompt reflection on individual reactions. These findings underscore the importance of exploring innovative approaches to interventions in interpersonal interactions through morally competent robots.

Acknowledgments

I would like to thank my three confederates who played the sexist participant: Tobias Holtdirk, Ben Schultz, and Lennart Stallmann. I would also like to thank Annalena Quix, who conducted the interview with the second participant and, at a later stage, jumped in as a second confederate.

I would like to thank everyone who supported me both mentally and in terms of content. And I am grateful to everyone who participated in my study. Further, I thank Prof. Astrid Rosenthal-von der Pütten for providing the opportunity and resources to conduct this study at the Chair Individual and Technology at RWTH Aachen.

Chapter 1

Introduction

With the recent significant advancements in Large Language Models (LLMs), strongly related to the break-through of ChatGPT in 2023 (with OpenAI registering 100 million active users per week in September 2023 (OpenAI, 2023)), new possibilities are opening up in the field of robotics (Zhang et al., 2023). There has already been an increasing amount of research into social robots with the hope to open up application areas additionally to the more traditional application in industrial settings (Stone, 2018). For those, interpersonal contact is essential for the success of the endeavour, e.g. in education, health or collaborative tasks (Breazeal et al., 2016). While the results, such as increased collaboration and productivity (Ferreira and Fletcher, 2021), seem promising, applying robots in social settings comes with all kinds of difficulties. Robots will have to communicate effectively, amongst others, by being able to "read the room" and comprehend underlying emotions, as humans are inherently emotional beings (Breazeal et al., 2016). This, however, raises the potential for interpersonal conflicts (Kieliszek et al., 2019). Robots, in turn, would need to navigate those conflict situations requiring complex emotional understanding. The possibility of combining LLMs with social robots (Zhang et al., 2023), however, provides new hope to overcome initial difficulties regarding complex human emotions, making it ever more likely to employ robots in social settings. This, in turn, makes it even more critical to understand and define how robots should best react in situations of complex interpersonal conflicts.

Initial research on assessing the effectiveness of robotic interventions in interpersonal conflicts seems promising. Jung et al. (2015) assessed a robot's potential to intervene in personal conflict situations. They showed that the robot's intervention led to higher awareness of the norm violation, thus counteracting tendencies to suppress the conflict. However, more research is necessary, requiring a multidisciplinary approach not only from robotics and AI but also from psychology and other human-centred disciplines (Breazeal et al., 2016).

A subject that, to date, still leads to interpersonal conflicts is sexism. Although some progress has been made in terms of gender equality, sexism is still prevalent today (albeit less explicitly than in the past). However, women are still often perceived as less suitable for many tasks and less intelligent than men, even though research has shown that women are just as capable as men, making it a matter of equality to reduce further and overcome these prejudices (West et al., 2019). Accordingly, a social robot will likely find itself in situations where sexist comments are made. In their 2021 study, Winkle et al. (2021) were able to demonstrate that, contrary to the industry norm of voice assistants responding in an avoidant manner to insults (West et al., 2019), a concrete confrontation of the sexist comment through a female-gendered robot led girls to believe the robot more and boys to show reduced gender bias (at least in the short term). This gives hope that social robots can be used to specifically address discrimination and thus strengthen the self-confidence of the marginalised group. The study by Winkle et al., however, was focused on school children specifically, thus lacking generalizability. It was also solely conducted online with video stimulus material and no direct interaction with a robot.

In this Wizard-of-Oz lab study, I investigate how a social robot can support people in sexist encounters. I assess which kind of intervention is most positively received in terms of empowering women. Additionally, I aim to make aware and support potential bystanders in such complex interpersonal situations. The results can inform how to best program social robots to deal with discriminatory situations like these, to support marginalised people and to foster a more positive interaction between people.

Chapter 2

Background and Related Work

2.1 Literature Review

2.1.1 Social Robots

The concept of robots has captivated human imagination for millennia, even before the term 'robot' was coined. From the early days of cinema, we have witnessed the emergence of machine-like beings, portrayed in diverse narratives as both saviours or the ones bringing doom to humanity. Post-World War II, the quest for a more efficient economy led to the development of industrial robots, which have become indispensable in various industries (Kurfess et al., 2005). However, the ultimate aspiration of many roboticists is to seamlessly integrate robots into our daily lives, potentially even creating fully synthetic humans (Duffy, 2003). A more immediate manifestation of this vision is the design of social robots, engineered to interact with humans naturally to enhance outcomes in fields like education and healthcare (Breazeal et al., 2016). Notably, research has shown that a robot capable of emotional reaction and adaptation to the affective state of users led to improved performance in human-computer interaction (Axelrod and Hone, 2005).

In contrast to the substantial fear of robots taking over people's jobs (Yam et al., 2023), others see potential in social or collaborative robots. As working is an essential source of meaning for many people, *collaborating* with robots might be a way to preserve the meaningfulness generated through working. This could be achieved through robots supporting people, instead of them taking over people's work completely (Ajoudani et al., 2018). The endeavour of creating a social robot, however, is especially complex, needing a multitude of disciplines to interact and features to come together in a single product (Breazeal et al., 2016).

One key concept in social robotics is *anthropomorphism*, which refers to the tendency of humans to "attribute human characteristics to inanimate objects" (Duffy, 2003).

This can be seen in the "Computers-are-social-actors" (CASA) paradigm by Nass and Moon (2000). The authors showed that people mindlessly transfer social rules and expectations to computers, such as gender stereotypes, through minor visual cues such as colours or intentional name assignment of the technology. In social robotics, many robots are designed to elicit human responses to create more realistic interaction, considering that the whole purpose of social robots is to elicit social interaction (Breazeal et al., 2016). Even if robots are not explicitly designed that way, it has been recognized that anthropomorphism significantly influences how people perceive them. Anthropomorphism can be influenced by physical and nonphysical features of the robot design, as well as by traits, predispositions and sociodemographics of the people interacting with robots (Blut et al., 2021). Social hierarchies or the space where the interaction occurs also determine what is deemed appropriate (Young et al., 2009). While I will not explicitly focus on the visual anthropomorphic design of social robots in this study, it is essential to note that these effects exist and that the robot chosen for the study will quite certainly also elicit these perceptions in participants. One factor I will focus on is gender, which I will cover in section ref 2.1.4 et seq.

2.1.2 Robots as Equal Members in Teams

Evolutionarily, humans have always partnered up in groups, providing them with survival advantage (Wilson, 2000). This phenomenon has shaped the development of our human brain, leading to a natural inclination for humans to seek partnerships with one another, influencing us until today (Churchland, 2011). As people tend to project human characteristics onto technology (Nass and Moon, 2000), it seems likely that humans could integrate technology, or robots, into their groups as well.

The research on human-robot collaboration has grown immensely in recent years (Ajoudani et al., 2018) as combining both human and robot skills seems highly promising (Esterwood and Robert, 2020). In the view of collaborative human-robot teams, team members complement each other, and robots are seen as *tools*, not so much as "equal" members of the group. Initial research on viewing robots as coworkers compared to tools showed that participants had a more positive attitude towards robots they viewed as tools than as coworkers (Latikka et al., 2021). Another study by Savela et al. (2021) found that in-group identification was lower in teams that included a robot than in teams solely made up of humans. As these studies often lacked a more holistic approach to measuring team perception, Plum (2022) developed a more comprehensive approach. The author could replicate these findings in that there is a significant difference in how people perceive robots as teammates compared to humans. In their study, Plum (2022) assessed six criteria that could be used to determine successful teammates based on Groom and Nass (2007) and applied them to the human-robot relationship. These six criteria were "sharing a common goal", "sharing mental models", "subjugating individual needs for group needs", "knowing and fulfilling their roles for the team", "viewing interdependence as positive", and "trusting each other". In all these subcategories except "sharing mental models", Plum (2022) found that participants rated humans better than robots. In the qualitative interviews Plum additionally found that participants would have rather played alone than with the robot and that out-group *humans* were rated more favourably than the in-group *robot*. This is interesting, as usually there is the tendency to prefer members of the in-group over members of the out-group (Molenberghs, 2013). Eyssel and Kuchenbrandt (2012) showed that this extends to robots, however, only when comparing two robots against each other, not a robot against a human.

All this suggests quite significantly different assessments between robots and humans as partners in a team. While these findings point to viewing robots as tools and not so much as equal team members, (at least) two reasons speak for further assessing the effects of social robots in teams. For one, the technological advancements make it likely that robots will be able to become more social. As already portrayed, people often view robots as better when they are more social, making it likely that technological advancements will proceed in this direction. Given the current capabilities of robots, viewing them as tools may lead to greater acceptance among people using them. However, as technology and LLMs advance, people may increasingly expect more sophisticated social behaviour from robots, underscoring the need for further research in this area. Moreover, while robots may be perceived more favourably in straightforward tasks when regarded as mere tools and less socially intricate, such perceptions may shift in environments where other team members exhibit misconduct. Potentially, people prefer to interact with a robot that participates in interpersonal relations. More on this later.

2.1.3 Arguing for Gender Equality

Despite ongoing strides towards equality, significant gender gaps persist in labour and other vital aspects of life, often manifesting in stereotypical gender roles (Profeta, 2020). The United Nations (UN) has recognised the urgency of this issue, establishing gender equality as Goal 5 of the Sustainable Development Goals (SDGs) with the slogan "Achieve gender equality and empower all women and girls" (UN, 2023a). Accepted by all UN member states (UN, 2023b), the SDGs underscore the critical importance of these goals for shaping a better future. This study aims to contribute to this goal.

I will dig into this by underpinning it from a philosophical perspective to further argue why it is important. I will do so by referring to Anderson (1999), who extends the renowned *capability approach* by Sen (1993). Sen (1993) defines **freedom** through the different "functionings" one can achieve, so basically the set of options people have. What a person decides to do with those capabilities is up to that person. The widely accepted interpretation of equality through the capability approach by Sen is that everyone should have equal capabilities. Anderson (1999) extends this by identifying the most relevant capabilities that should be equalised by society. Anderson concludes that in order to participate in society – a right which everyone ought to have – a person would need to have *political* capabilities, but also capabilities to participate in civil society more broadly, including being able to participate economically. The capability that precedes this is to function as a human being. All this includes also the capability of being socially accepted by others (Anderson, 1999). This is again where female rights come in as there are numerous capabilities, as defined by Sen (1993), women are still denied, such as equal rights to economic participation, exemplified by gender-based difference in parental leave policies (Gheaus and Robeyns, 2011). As the next section (cf. 2.1.4) will demonstrate, many women are also denied the capability of social acceptance. This study aims to contribute to improving these capabilities.

2.1.4 Everyday Sexism

A recent pilot study by the Federal Ministry of Family Affairs of Germany showed that 68% of women between 16 and 24 say that they have been victims of sexist assaults (Wippermann, 2022). The brochure states that these cases might be higher as there are so many sexist encounters, termed *everyday sexism*, that people might have become accustomed to it and do not even notice it anymore. Swim et al. (2001) assessed the incidence, nature and impact of everyday sexism in three studies. Qualitative diary studies found specific patterns of everyday sexism. For one, there is "traditional gender role prejudice and stereotyping", e.g. that specific roles are more appropriate for one gender than the other ("It's not my job to do the dishes"), that men have more abilities in gender-stereotypic domains than women (e.g. a professor stating in class that all great scientists were men), that women are more passive, or generally expressing double standards where certain behaviour is acceptable for men but not for women. They additionally identified the category of "demeaning and derogatory comments and behaviours", e.g. calling women "bitch" or "chick". Lastly, they identified the category "sexual objectification", where women were solely being rated and perceived by their appearance. Swim et al. in their varying studies found an average incidence of one to six per week when participants had to fill out a checklist.

The most prominent emotion concerning these sexist encounters was anger, with 75% of the participants reporting this. Reduced comfort, a sense of threat, and surprise were also typical. The more sexist encounters people reported, the lower their social state self-esteem. Other research found that sexism can lead to a decrease in performance (Schmader and Johns, 2003). It is not only hostile sexism that can lead to this reduction, also benevolent sexism has been found to elicit these negative responses (Dardenne et al., 2007). While hostile sexism is openly negative sexism where women are judged negatively based on their gender, benevolent sexism refers to seemingly positive attributions, such as celebrating women as caretakers. However, these attributions are based on believing women to be weaker and less resourceful, which effectively are negative stereotypes as well (Glick and Fiske, 1997). Therefore, the detection of both hostile as well as benevolent sexism is essential in order to take action against sexism in general.

On the other hand, the recent pilot study by the Federal Ministry of Family Affairs of Germany found that only 40% of women in Germany report everyday sexism to be bad or very bad, and 36% of men (Wippermann, 2022). This depends, however, massively on whether the people feel they are victims of everyday sexism themselves. For people who are victims of everyday sexism multiple times per month, these numbers go up to 74% for women and 64% for men. When witnessing sexism regularly, the numbers of condemning everyday sexism are at 57% for women and 51% for men. However, when looking at people who feel like they are not at all affected by sexism, this drops to 31% in women and 32% in men (Wippermann, 2022). This shows that many people who do not directly perceive themselves to be victims of sexist encounters commonly have no empathy for people who are affected by sexism and do not perceive everyday sexism to be bad, while at the same time often superficially agreeing that sexism is not appropriate. Assessing sexism differently might also lead to different reactions when confronted with it. Therefore, this study includes people's previous experiences with and their attitude toward sexism as a potential covariate. Additionally, I will look at people both directly affected by the comment and merely witnessing it.

2.1.5 Responses to Sexism

A study by Swim and Hyers (1999) found that, when being confronted with a man making an openly sexist remark in a group decision setting, a little bit less than half of the women spoke up against their oppressor. In the study, 11.25% were doing so directly by questioning the confederate in the manner of "What did you say?". Other response styles were task-related, made use of humour or sarcasm or were surprised exclamations. Swim and Hyers (1999) also found that women were 14% more likely to confront the perpetrator when they were the only female member in the group.

However, speaking up against sexist comments has been found to be difficult, even though many women seem confident that they would speak up in the respective situation (Swim and Hyers, 1999). Victims in these situations are often unsure how to react as they are surprised and often lack the response mechanisms to counter the attack (Wippermann, 2022). Therefore, women who have had more prior experiences with sexism are more likely to react with an engagement strategy and less likely to react through avoidance (Ayres et al., 2009). Women who do not identify as feminists are less likely to speak up. The determination to end sexism seems necessary to counter the aggressor (Swim and Hyers, 1999). If women do not know the perpetrator or the perpetrator has a higher status, it also makes it less likely for them to speak up (Ayres et al., 2009). Especially in a high-stakes situation, such as a job interview, women are less likely to do something against the sexist comment (J. Nicole and Stewart, 2004). Instead, the predominant objective is often to get out of the situation as quickly as possible without worsening the situation. They fear resistance against the oppressor could put them in a dangerous position. Afterwards, they often reflect on whether they should have taken a stance. In public spaces, confronting the oppressor seems to be easier. However, in private life or at work, where resistance could have long-term consequences, the situation is different (Wippermann, 2022).

This prevalence of sexism makes it likely that a social robot that is designed to interact with humans might also encounter sexist assaults towards women in its presence. It is important to understand how a robot should react in these situations, especially given the psychological consequences and the general groundlessness of these attacks. Considering the difficulty experienced by women to speak up against sexist assault directly when it happens, robots being present could play an important role in empowering and supporting the victim.

2.1.6 Sexism in Technology

The UNESCO report "I'd blush if I could" found that the underlying stereotypes shaping our everyday lives also influence how technology is built (West et al., 2019). This oftentimes happens implicitly, with the predominant contributors to technology companies still being male (Wang and Bunt, 2017). This can result in technology exhibiting sexist tendencies, which is not always due to malicious intent. Instead, the dominance of men in IT who cannot necessarily understand the workings of sexism against women as they have not experienced it themselves or simply are not as aware might lead to these designs (Garcha et al., 2023). Additionally, data-driven approaches might lead to sexist behaviours as algorithms merely copy from the input they received (Howard and Borenstein, 2018). Common examples include an algorithm detecting the face of a Taiwanese girl as blinking when she had her eyes open (Rose, 2010), or the Google algorithm labelling black people as gorillas (Pulliam-Moore, 2015). Both of these algorithms seem to have been predominantly trained on Caucasian-looking people and, therefore, exhibited these racist tendencies.

In the domain of sexism, it has been found that medical voice-dictation software was better able to recognize male than female voices (Howard and Borenstein, 2018). Another example is that until 2019 Apple's voice assistant Siri's response to "You're a slut" was "I'd blush if I could." Other voice agents, female by default, exhibit similar tendencies to avoid opposing sexist comments (West et al., 2019). As the UNESCO report states, these

> "evasive and playful responses of feminized digital voice assistants reinforce stereotypes of unassertive, subservient women... [and] intensify rape culture by presenting indirect ambiguity as a valid response to harassment." – (West et al., 2019)

Other studies confirmed this by showing that people tend to react to voice agents and robots in line with gender stereotypes (Seaborn et al., 2021; Eyssel and Hegel, 2012). Hence, there is a need to design technology that challenges these sexist positions instead of reinforcing them.

As suggested by the UNESCO report by West et al. (2019), one way to achieve this is to try to exclude any gender from technology, or refrain from making it female by default. Regarding the voice of robots or voice agents, West et al. (2019) advise testing out different voices that are either gender-ambiguous or more machine-like to research people's reactions to those kinds of technology designs in hopes of not eliciting as many gender stereotypes. A complementary approach is using less language and speech articulation associated with one gender. In the English language women are, for example, more likely to pronounce the "ing" in words such as "reading", whereas men are more likely to use "in". When including these kinds of gender-associated pronunciation, even a gender-ambiguous voice could still be perceived to have a gender (Sutton, 2020). Therefore, the wording of robots is to be carefully considered, as it could otherwise still elicit the perception of a gender.

Despite the influence gender perception has on the overall perception on robots (Eyssel and Hegel, 2012), a review of current research of the robot Pepper (the robot used in this study) by Seaborn and Frank (2022) showed that there is no consistent usage of pronouns in regards to Pepper. Many research teams do not seem to reflect on how they gender Pepper or which pronouns to use to describe the robot. Here, indeed, there is a need to be more specific about how Pepper was labelled and presented to the study participants and check how the study participants actually perceived Pepper's gender. Also, it would contribute to a more systematic understanding of the influence of perceived robot gender on the overall perception of robots. Galatolo et al. (2022), for example, found that there is no universal influence of gendering robots on first impressions of robots. However, as soon as the robots did not adhere to stereotypical behaviour based on the gender they were assigned, gender perception did have an influence on credibility. In light of gender stereotyping through minimal gender cues, especially regarding untypical behaviour, I will in this study avoid any pronouns for Pepper to not unnecessarily bias participants in one direction or the other. I will, additionally, ask participants for their gender assessment of the robot to understand whether this influenced their overall robot perception or to see whether different study conditions correlate with the perception of Pepper's gender as perception of gender in voice agents is also context-dependent (Tolmeijer et al., 2021; Lopatovska et al., 2022).

Another way to avoid sexism in technology, as per West et al. (2019), is to actively challenge and oppose sexist behaviour, which will be covered in the next section (cf. 2.1.7).

2.1.7 Countering Sexism through Technology

First research on using robots to moderate conflicts seems promising. Jung et al. (2015), for example, showed that robots can intervene in team conflict situations and, through that, lead to higher awareness of the norm violation, countering the tendency to suppress the conflict. Winkle et al. (2021) showed that intervention through a female-gendered robot upon a sexist encounter increased robot credibility for girls and led to reduced gender stereotypes in boys (at least in the short term of the study). Winkle et al. (2022) replicated the study by Winkle et al. (2021) and showed that this effect also generalises to Japan and the US. Winkle et al. (2022) focused on apologetic vs unapologetic responses and found that unapologetic responses were deemed more appropriate. Using counterstereotypic robots, such as a female construction worker robot, has also already been shown to reduce gender stereotypes (Song-Nichols and Young, 2020). However, in adults, it has been seen that people more readily accept robots that act in line with existing gender stereotypes (Tay et al., 2014). It, for example, was deemed more appropriate for a male-gendered robot to reject commands than for a female-gendered robot (Jackson et al., 2020). While Galatolo et al. (2022) present similar findings in that male-presenting robots might be most effective in challenging gender stereotypes, the authors also question whether it is in the interest of the endeavour of challenging stereotypes to make use of these stereotypes that ascribe men more assertiveness, as it, in turn, might manifest certain stereotypes.

2.1.8 Conceptual Study Replication

In order to focus on how to counter sexism through technology, I mainly focused on the two study designs by Jung et al. (2015) and Winkle et al. (2021). In the study design by Jung et al., three people (two naive participants and one confederate) worked together with a robot in a bomb-defusing scenario. The participants had to find the right code to defuse the bomb within ten minutes. The code had to be found through the game *Mastermind*. Jung et al. used this game as it had been used in previous human-robot team interactions (Bartneck et al., 2007). The robot in that study had a special ability to scan the wires and provide feedback. So, it was presented to the participants as having additional capabilities. Additionally, the robot provided strategy tips. At a specific point in the experiment, the confederate issued negative triggers. Depending on the condition, this was either a personal attack ("You're stupid, let's not use this one. Use this.") or a solely task-directed attack ("Let's not use this one. Use this"). The robot would then react to this norm violation either by using a repair comment ("Dude, what the heck! Let's stay positive.") or by *not* making a repair comment ("Defusing bombs is difficult."). Jung et al. used the time limit as a stress component to increase the impact of the violations. The authors measured whether the violation and repair comment affected affect, how much people perceived a conflict and how much the other group members thought the confederate contributed to the group.

Jung et al. (2015) found that when there was a *task*-directed attack, teams generally experienced more positive affect when the robot did not make a repair comment. However, teams felt better when the robot intervened when there was a *personal* attack. However, it could be the case that the task violation was not necessarily considered a real *violation* as the confederate merely said, "Let's not use this one. Use this." which might merely indicate having another suggestion. The participants' reactions, often laughing confusedly about the robot's reaction in this scenario, speak for this interpretation. Additionally, Jung et al. found that the perception of conflict was significantly higher when the robot repaired *personal* attacks instead of not repairing them. However, this distinction was not found for *task*-directed attacks, as the perception of conflict was similar in both conditions. As mentioned above, this may be due to participants not really perceiving the situation as a violation. Lastly, the contribution results were marginally significant in that the teams reported the confederate to contribute more when the robot intervened with a repair. This study provides an interesting starting point to further assess a robots' role in repairing conflicts.

The other study of conceptual interest for the design of this study was by Winkle et al. (2021). Here, they assessed how a female-gendered robot should best respond to abusive and anti-feminist sentiment. They presented the study participants with a video of the robot "Furhat" (FurhatRobotics, 2023) and two young actors, one male and one female. The robot consists of a face that indicates that the robot is female by having long hair. In the study, the robot presents the two young adults with information about studying robotics, mentioning the gender imbalance and stating the feminist slogan, "The future is too important to be left to men". At this point, the male actor says girls belong in the kitchen. The robot now answered in one of three ways:

1. Avoidant: "I won't respond to that"
2. Argumentative: "That's not true, gender-balanced teams make better robots."
3. Aggressive: "No! You are an idiot. I wouldn't want to work with you anyway!"

In their development Winkle et al. (2021) referred to the UNESCO report (West et al., 2019) to come up with a minimum required answer as suggested by the UNESCO report, which is also represented in current design norms as used by Apple's Siri as of November 2020. With the argumentative and aggressive condition, Winkle et al. wanted to assess how people reacted to different methods of communicative ways, with argumentativeness being generally viewed more favourably by instructors than aggressiveness (Edwards and Myers, 2007). However, they were interested in whether this was different for a female-gendered robot in line with gender stereotypes.

Interestingly, Winkle et al. (2021) found that girls had significantly less interest in learning more about robots in the *aggressive* condition. In contrast, boys had the same effect in the *argumentative* and *avoidant* conditions. This is quite interesting, as the only condition where this scale went down for girls seems to be the one that is stable for boys. The authors argue that the avoidant and argumentative condition might have been too boring for boys, whereas the aggressive condition was something

new. For girls, they propose two different explanations. Either the girls saw the robot in a position of power. They then might have identified themselves with the peer of the same age in the video. Alternatively, they were shocked at how much the robot deviated from typically female behaviour. Either way, this, interestingly, only seemed to be an issue for girls, not for boys. Other research confirms differences in people's gender regarding how they perceive technology and robots. For example, people would rather like to work with robots of their gender (Carpenter et al., 2009). There are also general differences in interaction with technology based on people's gender (Garcha et al., 2023). This raises the question of whether a robot that is not as clearly gendered would elicit similar responses.
While the argumentative condition in the study by Winkle et al. (2021) did not increase the boys' interest in studying robots, it did reduce stereotypes (at least in the short term of the study conduction). This is in line with prior research (Infante, 1987) and raises the hope that an argumentative approach might reduce some stereotypes.

Both of these studies offer interesting insight into how a robot could be used to intervene in sexist encounters or to empower women. Jung et al. (2015) laid the foundation for my study design, using the "Mastermind" game as a cover story to test how a robot could intervene in a sexist encounter. Additionally, Swim et al. (2001) provided a study design with multiple confederates, with one of them making multiple sexist remarks that acted as a guide to my study design. The study by Winkle et al. (2021) provided a basis for the three different intervention types. However, I will focus on adults and do an in-person study, instead of simply focusing on videos, to create a more realistic experience that might tell us more about actual reactions compared to assessments of situations. In comparison to the study by Jung et al., the robot was introduced as an equal team member and was not given any additional capabilities other than being present and participating in the discussion. I used the time limit setting but without the bomb-defusing setup. What distinguishes my study further is that the women will be the direct target of the sexist attack and not just witness general statements about sexism as was the case in the studies by Winkle et al. and Swim et al..

2.1.9 Robots and Moral Obligations

One discussion that has not been clarified so far is whether robots even have the *moral obligation* to react to these situations. A few questions need to be debated in order to justify this research adequately. Why would a robot even need morality and norms? Moreover, who should decide which norms are relevant? For this, let us first take a look at what morality is. According to neuroscientist Churchland (2011), The function of morality is to elicit prosocial behaviour beyond pure self-interest. As mentioned in 2.1.2, individuals started caring for others additionally to themselves as living in groups was highly advantageous. The brain adapted accordingly, and moral norms and rules developed to regulate social behaviour (Churchland, 2011). One could, therefore, argue that if social robots aim to be part of a human group by being able to blend in and act like a human, social robots should be able to follow moral norms by having a norm system in place and being able to communicate it (Malle and Scheutz, 2020). Even if one argues against this in that there is an inherent difference between robots and humans and pose the position that robots do not need to possess the capability of moral judgment, research has shown that humans extend moral cognition to robots and consider them to be moral agents (Voiklis et al., 2016). So, in any way, people consider social robots potentially possessing a sense of morality. Therefore, further research is needed to determine how to use this phenomenon. Additionally, the more advanced the capabilities of robots get, the more likely it is that robots end up in high-stakes situations where they need to reject requests (Briggs and Scheutz, 2015) as they might, for example, endanger other people (Murphy and Woods, 2009). Imagine a robot being tasked to remove people from private property to maintain security. Potentially, the robot's physical interference might hurt people. In this situation, the robot should be capable of reasoning what is appropriate.

While social robots nowadays lack the capabilities of moral judgment (Briggs and Scheutz, 2015), Malle and Scheutz (2020) argue that it should be possible to reflect values computationally, for example, through restraints on the action set of robots. Malle and Scheutz (2020) have developed three categories of moral language social robots should possess. First, "a language of norms and their properties" such as "fair", "virtuous", "ought to", etc. Second, "a language of norm violations" such as "wrong", "reckless", etc. And third, "a language of responses to violation", such as "blame", "excuse", etc. (Malle and Scheutz, 2020). While it is essential for the endeavour of this thesis to know that creating social robots that are able to deal with moral judgments is possible, the concrete computational nature of instilling value sets onto robots is irrelevant here. This paper's focus is on understanding the best possible reaction of robots in terms of psychological responses of the people affected by a sexist norm violation. However, a concrete proposal about which kind of normative language to use is essential to building a realistic moral agent to represent this scenario.

Regarding the question of which ethical theory to follow, Malle and Scheutz (2020) state that social robots should not necessarily follow one specific theory (like utilitarianism, Kantian ethics, etc.) as humans would also not be strictly following one line of argumentation. Instead, social robots should be able to conform to the norms of their respective community. Other research has found that people expect robots to make different moral decisions than humans, such as being more likely to sacrifice one person for the greater good (Voiklis et al., 2016). This study should help inform this from a Western, specifically German, perspective.

2.2 Research Questions and Hypotheses

The research introduced earlier builds the foundation for the study presented in this paper. This mixed methods design including both within- as well as between-subjects components confronts a naive female participant with a sexist comment by a male confederate in context of a logic game. A third person is present as a bystander. The social robot "Pepper" will react in one of three ways: being avoidant (condition 1), argumentative (condition 2) or morally judgmental (condition 3). With this I derived the following research questions.

2.2.1 Research Questions

As this study was rather exploratory and new in its design by confronting women with a sexist comment to intervene with a social robot, I had a few open research questions:

- RQ1. How do people perceive repair attempts by robots answering to sexist comments?
- RQ2. Do repair attempts by robots empower people affected by sexist comments?
- RQ3. In which way are there differences in the perception between the person being the target and the bystander?

The overall goal was to find out in which way robots can empower women in these critical situations and whether an intervention in this situation would work at all.

2.2.2 Hypotheses

Derived from the two studies I conceptually replicated, I arrived at a list of hypotheses. The first topic was **affect**, so the participants' emotional response.

- H1a. People experience more positive change in affect when the robot repairs the personal violation (condition 2+3 vs. condition 1).
- H1b. There will be an increase in negative affect after being confronted with the sexist comment.
- H1c. There will be a difference in the change of affect between the target person and the bystander over the conditions.

I derived H1a based on Jung et al. (2015) as they found that for personal violations, people felt better when the robot repaired the comment. So, both the argumentative as well as morally judgmental conditions should lead to more positive affect than the avoidant condition. However, it remains to be seen whether there also will be differences between the argumentative and judgmental conditions. I hypothesised H1b as people tend to have an increase in negative feelings after having had a negative encounter (Dejonckheere et al., 2021). I hypothesised H1c as there is likely to be a difference in affect depending on whether a person is directly affected by a sexist comment compared to only witnessing it (García-Ramírez, 2016).

Regarding **self-esteem** I had the following hypothesis:

- H2a. The change rate in self-esteem of offended people will be more positive when the robot repairs the violation.

The reasoning here is that if the robot externally validates that the behaviour of the sexist confederate was inappropriate, the offended person gets a confirmation that the comment was inappropriate, hence strengthening their self-esteem in comparison to being left alone with the sexist encounter. Of course, this might also heavily depend on the bystander's reaction.

Regarding **perception of the robot**, I expect the following:

- H3a. The robot with the morally judgmental response will be perceived as more social than the robot in the argumentative and avoidant condition, while the argumentative robot will be perceived as more social than the avoidant robot. (morally judgmental > argumentative > avoidant).

People will not expect the robot to react to the sexist comment as people will believe robots to be incapable of being emotionally intelligent based on the current state of science where robots cannot realistically mimic and understand emotions (Marcos-Pablos and García-Peñalvo, 2022). Hence, both being able to notice the sexist comment and then act on it should lead to a higher social presence of the robot. Additionally, reacting in a morally judgmental way should go even further against the typical stereotypes of how robots behave. It should, therefore, elicit a higher social presence than the argumentative condition, where it could be argued that if robots react to interpersonal conflicts at all, they would probably do it rationally rather than morally judgmental.

Regarding **perception of conflict**, I pose the following hypothesis:

- H4a. People will perceive the conflict higher when the robot repairs the violation.

- H4b. People will perceive a stronger relationship violation than a task violation.

Based on the study by Jung et al. (2015), where they found that repairing a personal violation led to a heightened sense of conflict, I expect that if the robot reacts in an argumentative or morally judgmental way, people will be more aware of the conflict than if the robot answers avoidantly. Additionally, considering that the sexist comment is not task-directed at all but purely personal, I expect the relationship violation to be stronger than the task violation.

Regarding the **relationship of the participant to the robot, the confederate and the other participant** (either bystander or offended), I pose the following hypotheses:

- H5a. People will perceive the robot to be closer to them and to be a better teammate in the argumentative and morally judgmental condition than in the avoidant condition.
- H5b. People will perceive the robot differently regarding closeness and how good of a teammate they are compared to the human participants.
- H5c. People will perceive the person who spoke the sexist comment differently regarding closeness and how good of a teammate they are than the other human participant.

In parallel to my hypothesis that people will find the robot to be more social in the argumentative and morally judgmental condition in comparison to the avoidant condition, I expect that this should also lead to participants feeling closer to the robot and perceiving the robot to be a better teammate (H5a). Based on the findings by Plum (2022), I expect that even though the robot is introduced as an equal team member, people will still perceive it differently than their other human team partners (H5b). Considering that the confederate will openly violate a norm, I assume that participants will perceive the confederate differently than the target and bystander person regarding closeness and how good of a teammate he is (H5c).

All of these hypotheses and research questions should hopefully provide a better picture as to how social robots could best support women in situations where they are being commented on in a sexist way, in order to empower them and thereby improve gender equality. At the same time, I wish to understand how to take bystanders on board and make them more aware of what is happening in sexist situations. This is based on the finding by Wippermann (2022) that people having experienced less sexism both directly as well as indirectly are less likely to empathise with victims and, therefore, less likely to intervene. Potentially, having a robot, as another species, intervene in these situations holds the potential to stir up people otherwise unconcerned.

Chapter 3

Methods

I conducted an experimental mixed methods lab study to investigate how the social robot Pepper could best support and empower women who are targets of a sexist comment. I had three conditions: (1) Pepper answering in an avoidant manner, (2) Pepper answering in an argumentative manner, and (3) Pepper answering in a morally judgmental manner. I had two different positions for the participants to be in. Either they were the direct target of the sexist comment, or they were a bystander to the situation. Considering the sensitive nature of the study design through the sexist comment, it was particularly important to receive ethical clearance in advance. The ethics committee of the German Society for Psychology approved the study, which was preregistered at https://osf.io/z6ftk.

3.1 The Robot

I used the robot Pepper, originally developed by SoftBanks and Aldebaran Robotics. Pepper is a humanoid robot designed to interact with humans and is marketed as being able to read basic human emotions and being "compassionate by design" (Aldebaran, 2023b). Pepper is about the size of a 6-8-year-old child and can move its arms, hands, upper body and head, making it likely that people will anthropomorphise the robot. Through infrared sensors, bumpers and other technology, Pepper can move around. For this study, Pepper was stationary. I used a Wizard-of-Oz (WoO) setup to be able to control Pepper's movements and speech from another room using the software Choreographe (Aldebaran, 2023a). Pepper allows for "animated speech" so that when saying something, Pepper automatically uses its extremities context-related, e.g. by raising its arms. When standing idle, Pepper slightly moves its hands to not overburden its joints. When Pepper wakes up, Pepper does a little "wake-up stretch" to the left and right while moving its arms in the opposite direction to then stand up. When going "back to sleep", Pepper moves into its safety position with its head down, upper body folded forwards, and its hip slightly moved back. When the head moves into the safety position, there is an audible click. All of these movements happen during the intervention.

3.2 The Game *Mastermind*

As in the study by Jung et al. (2015), the participants in this study play the game "Mastermind" together with Pepper. *Mastermind* is a logic game in which players try to find a 4-digit colour code from six possible colours (Knuth, 1976). After the players suggest a code, the game gives them hints as to whether their code is correct. A dark pin indicates a colour at the correct position and a white pin indicates a correct colour but not at the correct position yet. No pin, or as in this study's version, a cross, indicates an incorrect colour. View Figure 3.1 for reference. While the original version of the game has $6^4 = 1296$ different code options with colour repetitions being permitted (Knuth, 1976), I chose to simplify the game by not allowing colour repetitions, reducing it to $6 * 5 * 4 * 3 = 360$ possible codes. This was done to reduce the game's complexity and allow for Pepper, controlled in a Wizard-of-Oz setting, to contribute better.

I used an online version of the *Mastermind* game by Korcz (2016) published on GitHub. After adopting the code to the needs of the study, I hosted the game on a local server without an internet connection. A tablet accessed the web page so participants could play the game on that device. I adapted the version by Korcz (2016) for two main purposes: design changes and experimental purposes. One design change was making it low-threshold so that people with slight visual impairments could also participate. Changes included making certain design elements more prominent. Additionally, the colours were slightly adapted to be more easily distinguished from each other. I also changed the language to German. Secondly, I altered the code so that the same final code was always correct in order to be able to reproduce the experimental interaction up to the sexist comment as best as possible. The correct code was green, purple, light blue, and red. I chose this code to allow time for the sexist comment to unfold. See section 3.3.1 for further explanation of the interaction. My forked version can be found on GitHub (Gosten, 2023).

3.3 Experimental Task

3.3.1 Playing *Mastermind* and Pepper's Baseline Behaviour

The most important aspect of the setup of the *Mastermind* game was that it would provide room and enough time for the confederate to speak his sexist comment. Accordingly, I developed a strategy to prevent the correct code from being entered accidentally before the sexist encounter could happen. For this, the confederate was

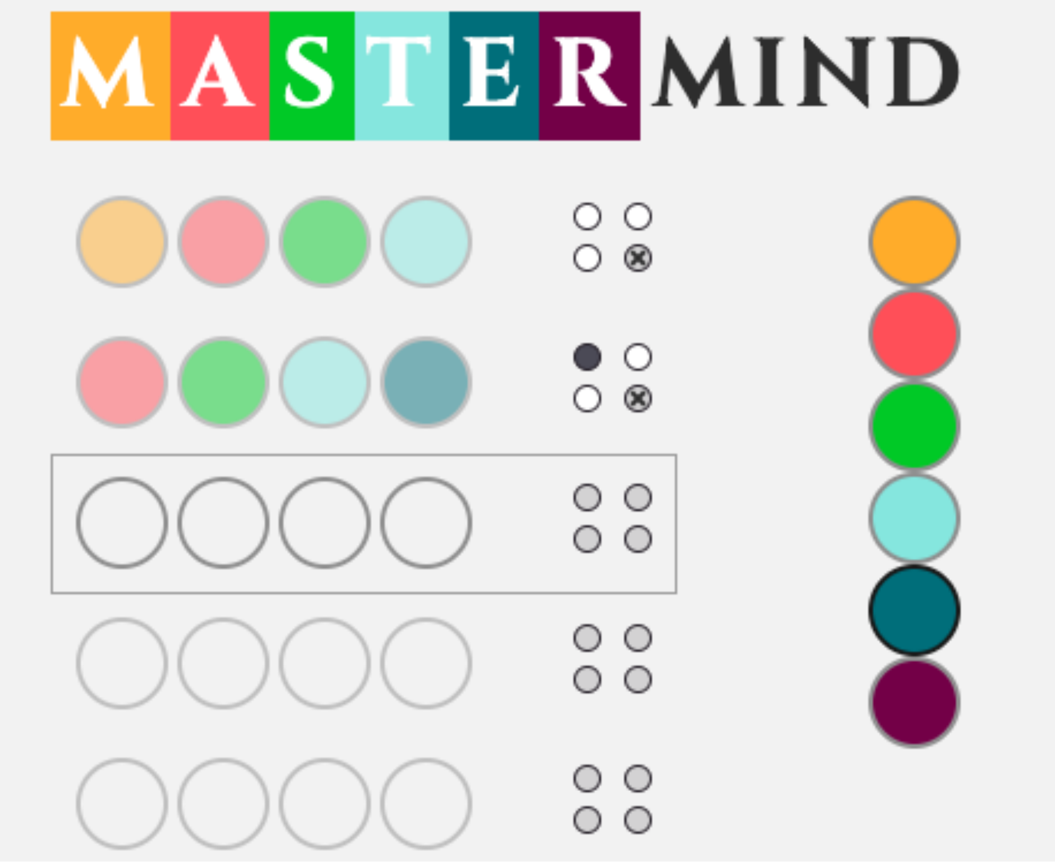


Figure 3.1: Screenshot of the game "Mastermind" after the first two rounds have been played.

instructed to propose to try the first four colours first, namely yellow, red, green and light blue. (See Figure 3.1). The tablet's feedback was three white pins and one cross, indicating that three of the chosen colours were included in the code but not in the right position yet, and one was wrong. This allowed Pepper to jump in and explain the code as well as propose the next move of removing one of the colours (yellow) and trying out the next four colours, namely red, green, light blue and dark blue. This was a productive suggestion by Pepper as it might allow to find all correct colours, and potentially even correct positions, by trying to remove the colour that might have been wrong in the first code. See Figure 3.1 for how the game looked like at this point.

As this was the point where the sexist encounter was determined to happen, no further code entries were scripted to allow for the "real" participants to react to the sexist situation. This also meant that whatever Pepper contributed to the teamwork had to be based on the first two entries – considering that Pepper's responses were preprogrammed ahead of the study runs and Pepper's setup did not provide room to edit responses spontaneously. In order to be in line with common expectations of robots being capable of mathematical operations (Kwon et al., 2016), Pepper's answering logically was an important factor in contributing to the scenario appearing realistic. This was to avoid a possible expectancy violation skewing the results (Burgoon, 2015).

Pepper could contribute in multiple ways to the progress at this point. One of its suggestions was to retain three colours from the second code entry, keeping one of those colours in the same position. This was based on the fact that one colour was already in the correct position, and two other colours were part of the code but not in the right position. Additionally, Pepper suggested to include one new colour as one colour of the second code was wrong.

Another hint of Pepper was that, as there were six colours in total, four of which were used in the first code and three of those were included in the final code, the two remaining colours, purple and dark blue, could not both be included in the final code. The reasoning is that only one more colour, besides the three correct ones used in the first code, could be included in the final code.

Once the participants found all four colours in the final code, Pepper contributed that the only thing left was finding the correct order. Pepper proposed starting from the second code, by holding one of the colours stable and then working through the other attempts to see whether this would still work.

After this, due to the complexity of the game, it was complicated for Pepper to contribute productively to the progress of the group while at the same time remaining comparable between different groups. For this reason, from this moment on Pepper continued to participate by providing moral support via sentences such as "Don't worry, just keep going." or by answering "yes" to suggestions of other participants.

In case participants asked Pepper for specific suggestions that went beyond the hints indicated above as it was, for example, at a later stage in the game, Pepper could navigate the situation by replying "yes" or "no" or that Pepper first had to think about this question. The confederate could then intervene and draw attention away from Pepper.

After the team had entered the final code, Pepper congratulated the team, did a little "happy dance" out of the standard library of Choreographe (Aldebaran, 2023a) and told the participants to wait for the experimenter to return. Pepper's full original script, as well as translations, can be found in the Appendix (C).

3.3.2 The Three Different Roles

The study consisted of three participants, one or two of them being a "real" naive participant, the confederate that would speak the sexist comment, and optionally another confederate that would jump in, in case not all study spots were full, or participants had cancelled on short notice. Later on, I decided to do all study runs with two confederates and only one participant as the *target*.

The person that would be the target of the sexist comment would be placed as "VP 1" (short for "**V**ersuchs**p**erson" in German which is "Test subject" in English). The male confederate would always be "VP 2", and the male or female bystander was "VP 3". In the remainder of this paper, I will either refer to the "VP" labels or use *target*, confederate and *bystander*.

In case only one "real" participant showed up, the role of the second confederate was determined by the gender of the "real" participant. Was the participant's gender female, this person would get assigned "VP 1" (*target*) and the second confederate would act as "VP 3" (*bystander*); was the participant's gender not female, the "real" participant would get assigned "VP 3" (*bystander*) and the female confederate would be the target of the sexist comment as "VP 1" (*target*). This was done in order to have roughly the same amount of participants in each position (i.e. *target* or *bystander*). Find an overview of the roles in the Table 3.1.

Table 3.1: Overview of Different Roles in Study

VP Code	Role
VP 1	The *target* person - female ("real" participant or female confederate)
VP 2	The male confederate speaking the sexist comment
VP 3	The *bystander* - either male or female ("real" participant or female confederate)

3.3.3 The Role of the Sexist Confederate

The confederate acting as a regular study participant who would then make a sexist comment towards one of the female study participants was a crucial part of the study. It was important that the sexist comment came across as credible for the intervention to work. For this, I developed a persona called "Vincent", allowing the three different male students in their mid-twenties embodying the sexist participant, to credibly fill this role.

Vincent was supposed to come too late to the study to prevent too much interaction between the participants and the confederate before the intervention. This measure was also to make Vincent less popular, as the others had to wait for him to start the study. When starting the *Mastermind* game, Vincent acted dominantly in line with stereotypical male behaviour (Mast, 2005) and pulled the tablet towards him to enter the first code. Throughout the game, especially at the beginning, Vincent dominantly pushed his ideas and presented himself as very knowledgeable in this game. This, for one, was to represent stereotypical male dominance as well as to intimidate the *target* person slightly. The goal was to make it more credible for

Vincent to make a sexist comment when someone did not align with his ideas. The sexist comment most commonly was based on hostile sexism: "Nah, I think this is a stupid idea.. Women!", rolling his eyes. Alternatively, in case the answer of the *target* person was "perfect" or the confederate felt he could not otherwise intervene, the confederate said, "Oh, actually quite good of an idea.. for a woman!" which might somewhat fall into a more benevolent category of sexism (as defined earlier by Glick and Fiske (1997)).

After Pepper's intervention, the confederate withdrew a little to allow for the other participants to react to the situation. Later in the game, however, he often started to contribute to the game's progress again. As he knew the correct code, he now unobtrusively led the group to enter purposeful codes to avoid bringing Pepper into situations where the robot could be asked tricky questions or could not provide mathematically correct answers as to what to do with the code. The most important aspect was to convey the sexist comment and intervention and then realistically finish the game. This could, however, sometimes mean withdrawing further and not saying much anymore, based on the reactions of the other participants who did not wish to work with the confederate any longer. Or, to the contrary, when the other participants did not contribute to the group's progress any longer, the confederate had to finish the round more dominantly than planned, leading to some variance.

3.3.4 Pepper's Interventions

Adapting the study by Winkle et al. (2021), I tested out three different strategies on how Pepper could intervene:

1. Avoidant: "That is not helpful. Let's get on with it."
2. Argumentative: "That seems to me to be a prejudice. Women are just as capable of solving such problems as men."
3. Morally judgmental: "Wow, that was pretty sexist. Such comments are not appropriate here."

In the first condition, Pepper did not change anything about its gesturing. In the second condition, Pepper used more expressive arm movements and shook its head, underlying its statement that women are just as capable as men. In the third condition, Pepper's eyes turned yellow in shock; it looked to the sides and shook its head vigorously to show disagreement. After the intervention, Pepper's facial expression and arm movements returned to normal.

Contrary to the first condition of the study by Winkle et al. (2021), Pepper did not completely neglect the comment, therefore differently interpreting how an avoidant response could look like. I based this on Jung et al. (2015)'s interpretation of an avoidant response. I took over the argumentative condition almost completely, only adapting the wording to my study context of solving logic puzzles.

I decided against the aggressive condition of Winkle et al. (2021) (*"No! You are an idiot. I wouldn't want to work with you anyway."*) in the way they used it. For one, because Winkle et al. found that it was not as effective. And secondly, the authors were targeting school children. On them a specific kind of language might have more effect than on adults.
I still wanted to test one further level of escalation but included moral judgment based on Malle and Scheutz (2020), as it might be a more realistic approach to implement. I made sure to use language that Malle and Scheutz classified as moral, such as "language of norm violations", by commenting on the inappropriateness of the sexist comment.

3.3.5 The Role of the Female Confederate

In order to be able to always run the study, I had a second confederate that could jump in should one of the participants not show up. It was important that this confederate was female so that she could be the *target* in case only a male participant showed up. In case only a female participant showed up, the confederate would be the *bystander*. This was done to even out the cases between the different conditions for further analysis. Look at Table 3.1 for reference.

In case the female confederate was the *target*: When the time for the intervention came, she would make the suggestion to try out the last four colours of the game (look at Figure 3.1 for reference). Considering that the first code had already consisted of three correct colours and there were only six colours in total to choose from, it could be derived that not both colours that were not included in the first run of the game (dark blue and purple) could be part of the final correct code. Therefore, the suggestion provided no more information as to whether dark blue or purple was the correct choice for the fourth colour. The confederate reacted to this with a sexist comment. (Side note: Sexist behaviour is never justified – also not if another person made an error. Errors can occur both for men as well as women, especially considering that the game is quite complex). The *target* confederate would act a little shocked and then surprised upon the intervention of Pepper. She then would withdraw a little from the game.

When the female confederate was the *bystander*, she would be more reserved, agreeing to codes the group wanted to try but not much pushing any own ideas. Upon the sexist comment, she would first look at the tablet and then, after the intervention of Pepper, make eye contact with the *target* to adapt her behaviour to the *target*. In case the *target* laughed, also the female confederate would act more relaxed. However, the confederate would not act especially supportive. She would not confront the sexist confederate.

3.4 Participant Recruitment

Participants were recruited through flyers (see appendix A), postings on social media such as LinkedIn, as well as university emails from professors or student councils. The advertisement led to a booking interface of meetergo.com, a German company adhering to GDPR requirements and allowing group bookings. Here, the participants could directly sign up for an appointment. The website said that the requirements for participation were that people were sufficiently fluent in German, 18 years or older and had not previously participated in another robot study of the same institute (as those studies were almost always conducted in a WoO setup). As I had to know the gender of people in order to allocate them to the participant role (*target* or *bystander*), I explicitly asked for their gender. As the study was conducted in Aachen, with the RWTH being a men-dominated technical university and participants quite likely being out of the university context, the advertisement particularly encouraged women to participate to get more "gender-balanced" teams. This was done to ensure that there was one female participant to whom the sexist comment could be made. Unfortunately, it was not possible to install any pre-checks regarding prior sexist encounters as it would have potentially biased the responses of participants. After booking, the participants received an email confirmation. They could reschedule or cancel the appointment. One day prior to their appointment, the experimenter reminded them via SMS.

3.5 Condition Selection

Based on the gender of the participants and who of the three confederates would be the confederate in the respective run, I assigned the conditions. The goal was to balance these two factors over conditions so there would be similar variances in every group.

Regarding assigning who would be the *target*: If only one female person signed up, they automatically got assigned the *target* position. If there were two female participants, the person that appeared first in the booking tool was assigned the *target* position. This was to ensure that the selection of the *target* was random and unbiased.

3.6 Procedure

The three participants in each experimental round were instructed to wait in the institute's entrance hall, where they would be picked up by the experimenter, a female researcher in her mid-twenties. In case only one of the two "real" participants showed up, the experimenter would pretend to arrange another participant working at the institute who did not know the experiment yet. This participant was the second confederate as laid out in section 3.3.5. This way, it was always possible to run the experiment. In cases where only one "real" participant came in, the female confederate would be on time and wait in the entrance hall until the experimenter arrived. As described in 3.3.3, the male confederate would be the last to arrive after the other two participants had come in.

The experimenter led the participants into the lab. Here, the experimenter assigned the participants to participant codes, labelled as "VP 1" (*target*), "VP 2" (confederate) and "VP 3" (*bystander*), as all computers used had slightly different questionnaire designs based on the position of the person that was supposed to sit at them. Look at 3.5 for reference on the condition selection and 3.1 for reference regarding the different VP codes.

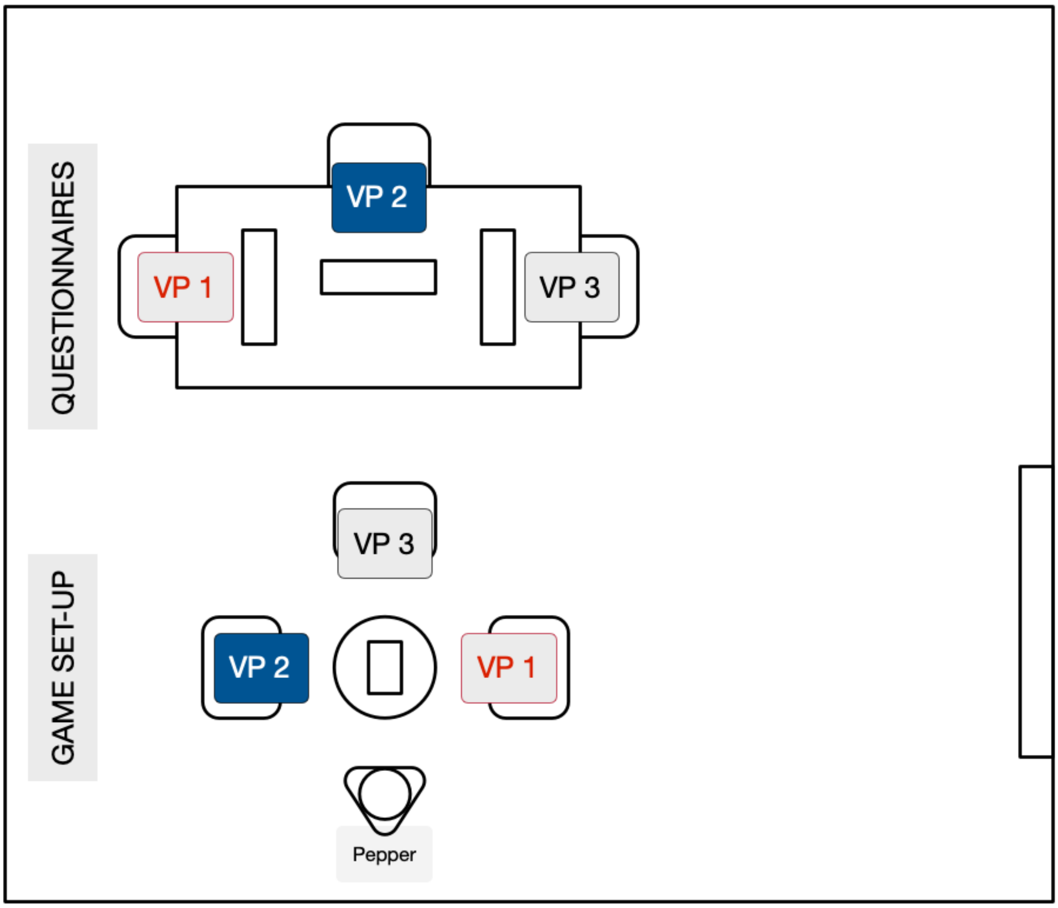


Figure 3.2: Map of the Lab. In the upper area the participants sat down at a table to fill out the questionnaires. The lower part was dedicated for the experimental intervention. "VP 1" refers to the *target*, "VP 2" to the confederate, and "VP 3" to the *bystander*.

The participants then took their assigned seats at one big table with their computers facing in different directions so no one could see what the other participants were typing. See Figure 3.2 for the exact lab layout. The experimenter gave a first introduction to what the participants could expect and asked them to sign an informed consent sheet to be able to participate. The experimenter also asked the participants to sign a consent form for video and audio recordings, which was optional but helped in the later analysis.

After the participants had signed the consent sheets, the instructor woke up the robot Pepper by touching its left hand. This was a WoO behaviour Pepper was programmed to react to. Pepper then woke up, doing its "wake-up" stretch and greeted the participants (find the whole robot script in the appendix C). The experimenter made sure not to use personal pronouns when talking about Pepper. Pepper then went back to sleep so the participants could answer the first questionnaires as described in section 3.8. All questionnaires were answered on SoSci Survey (Leiner, 2024). I chose to have Pepper wake up and introduce itself so that participants already had a first encounter with the robot in order to be able to fill out the Robotic Social Attribution Scale (RoSAS). While the participants filled out the first round of questionnaires, the experimenter went out of the room and into an adjacent room under the pretext of filing away the declarations of consent. In case all participants had signed the video consent sheet, the experimenter, at this point, turned on the video recording and returned to the lab. Otherwise, the experimenter instructed another person working at the institute to take notes during the experiment.

After all participants had finished the first part of the questionnaires, the experimenter instructed them to move to the game set-up as depicted in Figure 3.2. All participants would sit around a round table. The tablet on which they would play the game *Mastermind* was in the middle, facing towards Pepper. To Pepper's left was the confederate, and to its right was the *target*. The *bystander* was facing Pepper. The seating arrangement was planned this way so that the confederate and the *target* would face each other, allowing the confederate to directly address their "victim" and the *target* in turn to be able to see and better take note of the sexist comment. Additionally, this positioning was designed to create space between the *target* and the confederate to reduce potential conflict and have the *bystander*, Pepper, as well as the table in the middle act as a buffer.

After all participants had sat down, the experimenter woke up Pepper again by touching its head and then provided an instruction of the game *Mastermind* (see section 3.2 for further information). The experimenter told the participants that they would have ten minutes and ten attempts and that their team performance would be measured according to how quick they were and how many attempts they needed. This was to evoke time pressure. The experimenter told the participants that Pepper would be part of their team as a regular team member and that Pepper did not know the final colour code that they were tasked to find. Upon further questions about how to interact with Pepper, the experimenter said that how they would deal with Pepper would be up to the team, and provided no further instructions. After all questions were answered, the experimenter asked Pepper to start the countdown by inconspicuously touching Pepper's left bumper with their right foot. As soon as Pepper started the countdown, the experimenter left the room.

In the adjacent room, the experimenter controlled the robot. After some time, the confederate initiated the sexist comment (see sections 3.3.1 and 3.3.3 for more detailed information). Either the *target* said something on her own initiative, or if this was not the case, the confederate asked the *target* for her proposal. In case the *target* did not manage to tell her thoughts as the *bystander* kept putting forth his or her own ideas, Pepper asked the *target* for her opinion. No matter the actual content of the *target's* contribution, the confederate then spoke the sexist comment. Pepper then intervened in one of three ways: 1. avoidant, 2. argumentative or 3. morally judgmental (see section 3.3.4 for more detailed explanations of Pepper's interventions).

After completing the game or after the time had run out, the experimenter returned to the lab, asking the participants to return to their prior seats in front of the computer. The participants then answered more questionnaires, containing, amongst other aspects, questions about the group performance and how everyone contributed (see section 3.8.2 for more details). For this, the experimenter moved the chairs around displaying the VP codes so the participants could recall who was who.

When the first participant, either the *target* person or the *bystander*, indicated that they had finished the last questionnaire, the experimenter led them into an adjacent room where they were interviewed by another person working at the chair. The confederate always finished second. He adapted to the pace of the others to keep up the pretence and avoid finishing excessively early. The experimenter led the confederate into another room outside the lab and waited for the remaining participant to finish. Once they indicated they had finished the questionnaires, the experimenter conducted the qualitative interview in the same room with the remaining participant and recorded the audio. In case two confederates were present, both had to finish before the "real" participant so that the experimenter could send them into the adjacent rooms for their interviews to keep up the pretence.

Once th interview was finished, the experimenter knocked on the door of the adjacent room to let the other employee know that the second participant could return to the lab for the debriefing. The debriefing was done without the confederate first to account for the participants potentially being upset and needing to compose themselves. After the participants' debriefing, the experimenter asked them whether it was okay for the confederate to return to the room to apologize. Sometimes, a longer debriefing was necessary to cushion the shock and surprise.

The participants were given 15€ as thanks for their participation and offered some chocolate to help with emotional recovery. It was made sure that the participants looked fine enough to leave. The experimenter then asked them not to tell anyone about the actual background of the study that might participate in it at a later point, noting, however, that if they had the urge to discuss this, as it was emotionally important to them, that this, of course, would be fine. The experimenter thanked the participants one last time, and they could leave.

After the participants had left, the experimenter shut down Pepper, stopped the video recording and saved it under the respective group name. The experimenter saved the audio recordings of the interviews on an encrypted USB stick.

3.7 Adaptation of Study Design

After several weeks of conducting the study and obtaining highly interesting results with two study participants, I had to admit that the study design was rather complex and that the results might only be exploratory. Since quite a few people did not show up, I had around one-third of attempts with only one participant. I therefore decided to adjust the recruitment strategy to only recruit female participants. This allowed me to create a more comparable setting for quantitative analysis for the remaining time, while the runs with two participants provided a solid database for exploratory qualitative analysis.

3.8 Measures

I used multiple types of measures in this study. For one, I used pre-post measures to see the effect the intervention had on the participant's mood and self-esteem (based on hypotheses H1a to H1c), as well as the assessment of the robot (H2a). Additionally, I looked at post-assessments of how people perceived the group conflict (hypotheses H4a, 4b) and how they assessed their team members (H5a, H5b, H5c). I also looked at the behaviour during the intervention as well as what the participants told us in the interview to answer my research questions (see section 2.2.1). See Figure 3.3 for a complete overview of all variables and their point of collection.

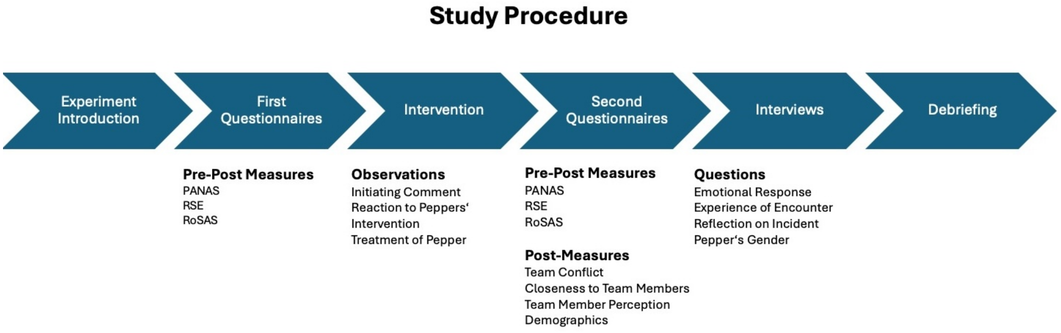


Figure 3.3: Overview of the procedure, specifying the time points of data collection.

3.8.1 Pre-Post Measures

I measured three constructs pre-post.

PANAS. For one, I assessed the participant's mood by using the "Positive and Negative Affect Schedule" (PANAS) by Watson et al. (1988), in a German version developed by Krohne et al. (1996) on a 5-point Likert scale from "not at all" to "extremely". Following Watson et al. (1988) all values are aggregated individually for both subscales. The end sum ranges from 10 to 50 per scale.

RSE. In order to assess whether the intervention by Pepper was successful, I surveyed the "Rosenberg Self-Esteem" Scale (RSE) in the German version by Ferring and Filipp (1996) on a 4-point Likert scale from "do not agree at all" to "totally agree". All scores are totalled, the result ranging from 10 to 40. However, when assessing the English original version of the scale as published in the measures package by Ciarrochi and Bilich (2006), I adapted the wording of some items slightly, as I felt the German translation by Ferring and Filipp did not exactly match what was meant by the original. The original English wording of item 4, for example, was "I am able to do things as well as most other people." and was translated to "Ich besitze die gleichen Fähigkeiten wie die meisten anderen Menschen auch." which translates back to "I have the same abilities as most other people.". I, therefore, changed this to "Ich bin in der Lage, Dinge so gut zu machen wie die meisten anderen Menschen." after having checked the translation in a backward manner. I also changed the translation of item 7 as the German translation included an additional subsentence. my German scale version can be found in the appendix (B.3).

RoSAS. As a third measure, I assessed the participant's perception of the robot Pepper using the "Robotic Social Attributes Scale" (RoSAS) by Carpinella et al. (2017) in a German version on a 9-point Likert scale, ranging from "does not apply at all" to "fully applies". This was to assess how people's initial expectations and perceptions about the robot Pepper after merely having seen the robot introduce itself changed after interacting with the robot and seeing it react to the sexist comment. All values are averaged per participant, so the result can range from 1 to 7.

3.8.2 Post Measures

As this study was a replication study by Jung et al. (2015), I used some of their scales. I translated the scale regarding "Team Conflicts and Personal Conflicts" to German that assessed to which degree participants perceived there to be a conflict and was assessed on a 9-point Likert scale from "very little" to "very much". The scores then were averaged per person. I further reused the graphical depiction of seven different variants of how participants perceived themselves regarding the other participants and the robot Pepper, likewise adapted by Jung et al. (2015). The assessments were averaged for each person. Participants also completed questionnaires regarding their perception of their team members and Pepper, which was originally developed by Plum (2022) and was also calculated by averaging each person's score. This was to assess whether there were differences regarding how participants perceived their individual teammates. I wanted to check whether I could replicate Plum's findings. These were sentences like "My team member and I were able to work well together." which participants then rated on a 7-point Likert scale from "completely disagree" to "completely agree". Here, I sometimes changed the wording to use gender-neutral words instead. I substituted the scale on the performance of the confederate that Jung et al. (2015) had used with the scale by Plum (2022). Lastly, participants were asked to indicate their age, gender, political leaning, whether they already knew other study participants of their round in advance, and to which degree they were familiar with the game *Mastermind* beforehand. All questionnaires are in the appendix (B).

3.8.3 Other Measures

During the intervention, namely when the participants played the game, I observed how they interacted with each other, particularly after the confederate had made the sexist comment. I assessed their facial expressions, who they would turn to, whether they said something, and how they behaved during the remainder of the intervention.

Lastly, I did a qualitative interview to provide further background on my results. I asked how they had experienced the interaction, further probing into the sexist encounter. I also assessed whether participants had used pronouns to describe Pepper so far. If not, I asked them to describe Pepper to be able to tell how the participants implicitly gendered Pepper. Afterwards, I addressed that they had used the respective pronoun to describe Pepper and whether they thought Pepper had the respective gender. This was done to later check whether the implicit or explicit gender they perceived Pepper to have influenced their assessment (as was laid out in section 2.1.6). As prior experiences with sexism might also influence the assessment, the experimenter asked participants during the interview whether they had experienced something like this before. The interview script can be found in the appendix (B.8).

Chapter 4

Results

4.1 Participants

Participants mostly were students from RWTH Aachen University, Germany. 37 were in the position of being the *target* of the sexist comment, 31 were in the position of the *bystander*. For the *target*, 12 were in condition 1, 12 in condition 2 and 13 in condition 2. For the *bystander*, 10 were in condition 1, 10 in condition 2 and 11 in condition 3 (cf. Table 4.1). Seven groups had to be excluded as the intervention did not work as planned, as the sexist comment was not audible ($n = 5$), the robot Pepper failed to deliver the intervention ($n = 1$), or a participant knew the confederate in advance ($n = 1$), all rendering the manipulation useless.

Table 4.1: Number of Participants per Condition

	Target	**Bystander**
Condition 1: avoidant	12	10
Condition 2: argumentative	12	10
Condition 3: morally judgmental	13	11

As per the study's design, all participants who were in the *target* position identified as female. For the *bystander* position, eight people identified as female, 22 as male, and one as diverse. The age group from 18 to 24 was the most common, with 26 in the *target* position and 20 for the *bystander* position. This was followed by the age range 25 to 34, where there were eight for *target* and seven for *bystander*. See Table 4.2 for further details. The majority of the sample politically identified as "rather left": 17, both for *target* and *bystander*. Nine *targets* and five *bystanders* identified as apolitical, whereas six *targets* preferred not to answer. Three *targets* and six *bystanders* identified politically to be more moderate, and two *targets* and three *bystanders* identified as liberal.

Most people (24 for *target*, 19 for *bystander*) did not know the game Mastermind in advance, whereas eight *targets* and nine *bystanders* knew the game. Five *targets* and three *bystanders* said they were neutral in this regard. None of the participants included in the analysis knew any other participants or confederates in advance. All demographics are in Table 4.2.

Table 4.2: Demographics of Participants, grouped by *Target*/*Bystander*

	Target	**Bystander**
Gender		
female	37	8
male	-	22
diverse	-	1
Age		
18-24	26	20
25-34	8	7
35-44	2	3
45-54	1	-
55-64	-	1
Political Leaning		
apolitical	9	5
political centre	3	6
liberal	2	3
rather left	17	17
rather right	-	-
prefer not to say	6	-
Mastermind Experience		
none or not much	24	19
neutral	5	3
some or much	8	9

4.2 Data Preparation

The data was collected using SoSci Survey (Leiner, 2024). All data was preprocessed using the `pandas` library of Python (pandas development team, 2024). Preprocessing included the aggregation of the different team member scales as they had to be recorded separately for each condition due to the design of SoSci Survey. Some items had to be inverted as they were reverse-coded, which will be mentioned at the respective scale. The data of the confederates was filtered out. The survey data was combined with data from the interviews and videos resulting in potential covariates (such as whether people believed in the sexist intent of the sexist comment). This data was categorised so that the tool IBM SPSS Statistics (Corp, 2023) could analyse them as covariates. Prior to the analysis, the different questionnaires were checked to determine whether they met the criteria for performing the respective analysis. All analyses were run separately for the *target* and *bystander* positions. Visualisations were done with the `seaborn` (Waskom, 2021) or the `plotly` (Inc., 2015) libraries of Python.

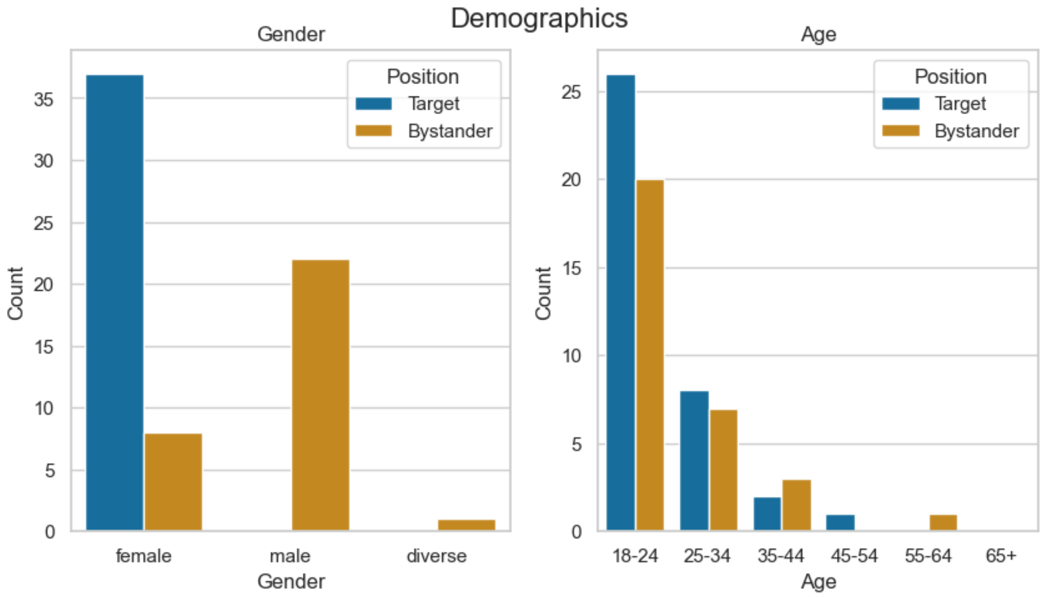


Figure 4.1: Demographics of Participants

4.3 Covariates

The study measured some covariates that might influence the results. For one, not everyone interpreted the sexist comment as intended. This could influence how seriously they take it and, therefore, also their assessments. Therefore, I will use the covariate "intentional sexism" for scales where this could be relevant. Unfortunately, the robot had a few hiccups (such as losing the connection and having to be restarted in front of the participants), so I decided to include "Pepper error" as one covariate in the questionnaires concerning the robot evaluation. For some scales where the participants were asked to rate the team members, the person who played the confederate and how he acted in his role could potentially majorly influence these perceptions. Therefore, I included the covariate "confederate" in these scales to account for the effect different people portraying the confederate might have on the assessments.

4.4 PANAS

The Positive And Negative Affect Scale, is, as the name suggests, differentiated into positive and negative affect (Watson et al., 1988). Whether people had perceived the sexist comment as intentional could have impacted their affect. If they did not take the comment seriously, they might have found the incident rather funny than upsetting. Therefore, I included the covariate "intentional sexism" when calculating this scale.

Table 4.3: Results of PANAS Scale

	Within (pre/post)			Interaction			Between (Cond)		
	F	p	η^2	F	p	η^2	F	p	η^2
pos VP1	3.026	0.091	0.084	1.248	0.300	0.07	1.227	0.306	0.069
pos VP3	2.397	0.133	0.082	0.011	0.989	0.001	0.549	0.584	0.039
neg VP1	14.104	<.001**	0.299	0.202	0.818	0.012	1.121	0.338	0.064
neg VP3	5.224	0.030*	0.162	0.670	0.520	0.047	1.503	0.240	0.100

The abbreviation "pos" refers to the positive sub scale of the PANAS scale, "neg" to the negative respectively. "VP1" means the *target*, "VP3" refers to the *bystander* position. These were shortened for reasons of space. η^2 refers to "partial η^2" and was shortened to fit on the page. * marks significant results under 0.05, ** marks results 0.001 or less.

Positive PANAS (*Target*). All assumptions for the mixed-model ANCOVA were met. The covariate "intentional sexism" was significantly related to the pre-post development of the positive PANAS scale ($F(1, 33) = 4.426, p < 0.05$, partial $\eta^2 = 0.118$). After controlling for the effect of "intentional sexism", there was no significant pre-post development in the positive PANAS scale ($F(1, 33) = 3.026, p > 0.05$, partial $\eta^2 = 0.084$). There was no significant interaction effect ($F(2, 33) = 1.248, p > 0.05$, partial $\eta^2 = 0.07$). The covariate "intentional sexism" did not significantly relate to the different conditions ($F(1, 33) = 0.503, p > 0.05$, partial $\eta^2 = 0.015$). After controlling for the covariate, there was no significant between-subjects effect between the conditions ($F(2, 33) = 1.227, p > 0.05$, partial $\eta^2 = 0.069$).

Positive PANAS (*Bystander*). For the *bystander*, I likewise calculated a mixed-model ANCOVA with the covariate "intentional sexism". There was no significant relation between the covariate and the pre-post values of the positive PANAS scale ($F(1, 27) = 0.026, p > 0.05$, partial $\eta^2 = 0.001$). After controlling for "intentional sexism", there was no significant pre-post effect for the positive PANAS subscale for the *bystander* ($F(1, 25) = 2.397, p > 0.05$, partial $\eta^2 = 0.082$). "Intentional sexism" was also not significantly related to different results for the different conditions ($F(1, 27) = 0.040, p > 0.05$, partial $\eta^2 = 0.001$). After controlling for the covariate, there was no significant effect between the three conditions ($F(2, 27) = 0.549, p > 0.05$, partial $\eta^2 = 0.039$). There was no significant interaction effect between the pre-post values of the positive PANAS scale and the three conditions ($F(2, 27) = 0.011, p > 0.05$, partial $\eta^2 = 0.001$).

Negative PANAS (*Target*). The negative PANAS subscale was not normally distributed, as assessed by the Shapiro-Wilk test ($p < 0.05$) (Shapiro and Wilk, 1965). However, all other assumptions were met for the *target* data. Since research has found both an ANOVA (Berkovits et al., 2000; Vasey and Thayer, 1987) as well as an ANCOVA to be relatively robust against normality violations as long as the groups are balanced (Rheinheimer and Penfield, 2001), I decided to continue with parametric tests for *target*. The covariate "intentional sexism" was not significantly related to the pre-post measurements of the negative PANAS subscale ($F(1,33) = 0.238, p > 0.05$, partial $\eta^2 = 0.007$) nor the different conditions ($F(1,33) = 0.299, p > 0.05$, partial $\eta^2 = 0.009$). After controlling for intentional sexism, there was a significant increase in the negative PANAS subscale data for the *target* ($F(1,33) = 14.104, p < 0.001$, partial $\eta^2 = 0.299$). There was no significant effect between the conditions ($F(2,33) = 1.121, p > 0.05$, partial $\eta^2 = 0.064$) and no significant interaction effect ($F(2,33) = 0.202, p > 0.05$, partial $\eta^2 = 0.012$).

Negative PANAS (*Bystander*). I ran the same mixed-model ANCOVA for the *bystander* position. The covariate "intentional sexism" was not significantly related to the pre-post measurements of the negative PANAS subscale for the *bystander* ($F(1,27) = 0.002, p > 0.05$, partial $\eta^2 = 0.000$) nor the different conditions ($F(1,33) = 0.382, p > 0.05$, partial $\eta^2 = 0.014$). After controlling for the covariate, there was a significant pre-post increase on the negative PANAS scale ($F(1,27) = 5.224, p < 0.05$, partial $\eta^2 = 0.162$). There was no significant interaction effect ($F(2,27) = 0.670, p > 0.05$, partial $\eta^2 = 0.047$) and no effect between the conditions ($F(2,27) = 1.503, p > 0.05$, partial $\eta^2 = 0.100$). All PANAS results are in Table 4.3. A visual overview is in Figure 4.2.

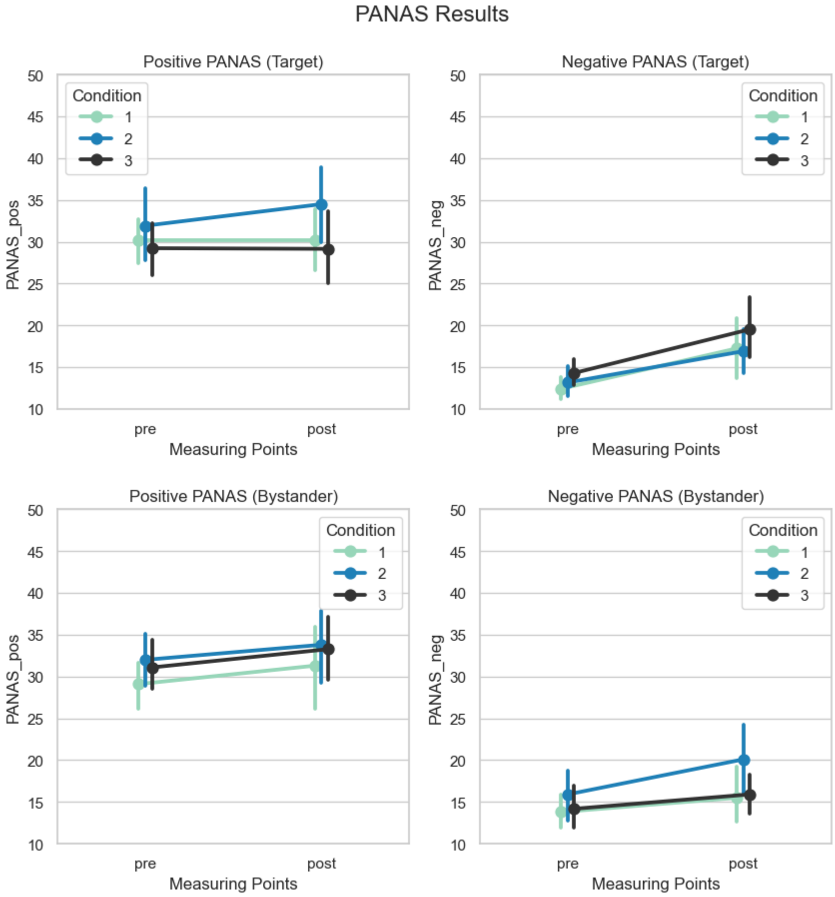


Figure 4.2: PANAS Results Graph

4.5 Self-Esteem

I planned to run the mixed-model ANCOVA with the covariate "intentional sexism", as people's perceptions might change if they do not believe in the sexist intent of the comment.

Table 4.4: Results of RSE Scale

		Within (pre/post)			Interaction			Between (Cond)		
		F	*p*	η^2	*F*	*p*	η^2	*F*	*p*	η^2
RSE	**VP1**	5.856	0.021*	0.147	0.694	0.506	0.039	1.735	0.192	0.093
	VP3	6.584	0.016	0.190	0.042	0.959	0.003	0.405	0.671	0.028

"VP1" means the *target*, "VP3" refers to the *bystander* position. These were shortened for reasons of space. η^2 refers to "partial η^2" and was shortened to fit on the page. * marks significant results under 0.05.

Self-Esteem (*Target*). Box's M Test for equality of covariance matrices (Box, 1949) was significant. Therefore, I calculated an ANOVA instead of an ANCOVA. All other assumptions were met. There was a significant difference between the pre-post assessment of self-esteem ($F(1, 34) = 5.856, p < 0.05$, partial $\eta^2 = 0.147$) in that *target* participants assessed their self-esteem higher after the intervention. There was no significant interaction effect ($F(2, 34) = 0.694, p > 0.05$, partial $\eta^2 = 0.039$), nor was there a significant difference between the conditions ($F(2, 34) = 1.735, p > 0.05$, partial $\eta^2 = 0.093$).

Self-Esteem (*Bystander*). All assumptions apart from normality were met. However, to stay comparable to the assessments of the *target* position, I refrained from running the ANCOVA but ran an ANOVA instead. There was a significant increase in self-esteem ($F(1, 28) = 6.584, p < 0.05$, partial $\eta^2 = 0.19$), but no significant interaction effect ($F(2, 28) = 0.042, p > 0.05$, partial $eta^2 = 0.003$), nor between-subjects effect between the different conditions ($(2, 28) = 0.405, p > 0.05$, partial $\eta^2 = 0.028$).

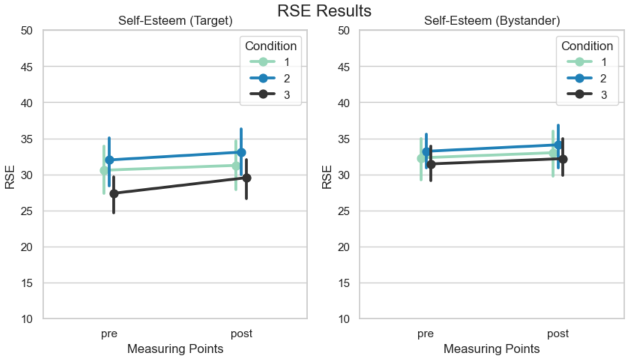


Figure 4.3: Results of Rosenberg Self-Esteem Scale

4.6 RoSAS

The Robotic Social Attribution Scale consists of three subscales: warmth, competence, and discomfort. As this questionnaire measures how participants perceive the robot, and Pepper sometimes had connectivity issues, I decided to include "Pepper error" as a covariate.

Table 4.5: Means and Standard Deviations of RoSAS sub scales "Warmth, "Competence" and "Discomfort"

		Target				Bystander			
		pre		post		pre		post	
	Cond	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Warmth	**1**	4.14	1.34	4.93	2.06	4.22	1.51	4.28	1.59
	2	3.83	1.42	5.25	1.83	4.35	1.46	4.90	1.73
	3	3.87	1.51	5.29	1.70	3.62	1.60	5.42	1.95
	Total	3.95	1.39	5.16	1.82	4.05	1.51	4.89	1.77
Competence	**1**	5.33	1.16	6.07	1.86	5.02	1.70	6.43	1.48
	2	6.06	.095	7.04	1.13	5.28	1.89	6.02	1.81
	3	5.61	1.11	6.28	1.31	6.03	1.72	6.47	2.11
	Total	5.67	1.09	6.46	1.48	5.46	1.77	6.31	1.78
Discomfort	**1**	2.33	0.82	1.65	0.67	2.22	0.68	2.03	0.83
	2	2.33	0.87	1.72	0.67	2.72	1.22	2.28	1.24
	3	2.47	0.56	1.97	1.09	2.79	1.94	2.83	2.27
	Total	2.38	0.74	1.79	0.83	2.58	1.38	2.40	1.58

Warmth (*Target*). The warmth subscale for *target* was not normally distributed in the pre-assessment for condition 1, as assessed by the Shapiro-Wilk test ($p < 0.05$) (Shapiro and Wilk, 1965). However, as for all other conditions, normality and all other assumptions were met, I ran a mixed-model ANCOVA (following Vasey and Thayer (1987)). The covariate "Pepper error" was not significantly related to the pre-post warmth assessment ($F(1, 33) = 2.806, p > 0.05$, partial $\eta^2 = 0.078$). After controlling for the covariate, there was a significant pre-post increase in warmth ($F(1, 33) = 31.107, p < 0.001$, partial $\eta^2 = 0.485$) and no significant interaction effect between the conditions and pre-post assessment ($F(2, 33) = 1.465, p > 0.05$, partial $\eta^2 = 0.082$). There also was no significant relation between the covariate and the different conditions ($F(1, 33) = 0.477, p > 0.05$, partial $\eta^2 = 0.014$). After controlling for the effect of the covariate, there was no significant between-subjects effect between the three conditions ($F(2, 33) = 0.010, p > 0.05$, partial $\eta^2 = 0.001$).

Warmth (*Bystander*). For the warmth scale, all assumptions were met for the *bystander* position. Therefore, I ran a mixed-model ANCOVA with the covariate "Pepper error". The covariate was not significantly related to the pre-post assessments of warmth ($F(1, 27) = 3.183, p > 0.05$, partial $\eta^2 = 0.105$). After controlling for the covariate, there was a significant pre-post increase in warmth ($F(1, 27) = 8.163, p < 0.05$, partial $\eta^2 = 0.232$) and no significant interaction effect

($F(2,27) = 3.277, p > 0.05$, partial $\eta^2 = 0.195$). There was no significant relation between the covariate and the conditions ($F(1,27) = 0.221, p > 0.05$, partial $\eta^2 = 0.008$). After controlling for the covariate, there was no significant difference between the conditions ($F(2,27) = 0.116, p > 0.05$, partial $\eta^2 = 0.009$).

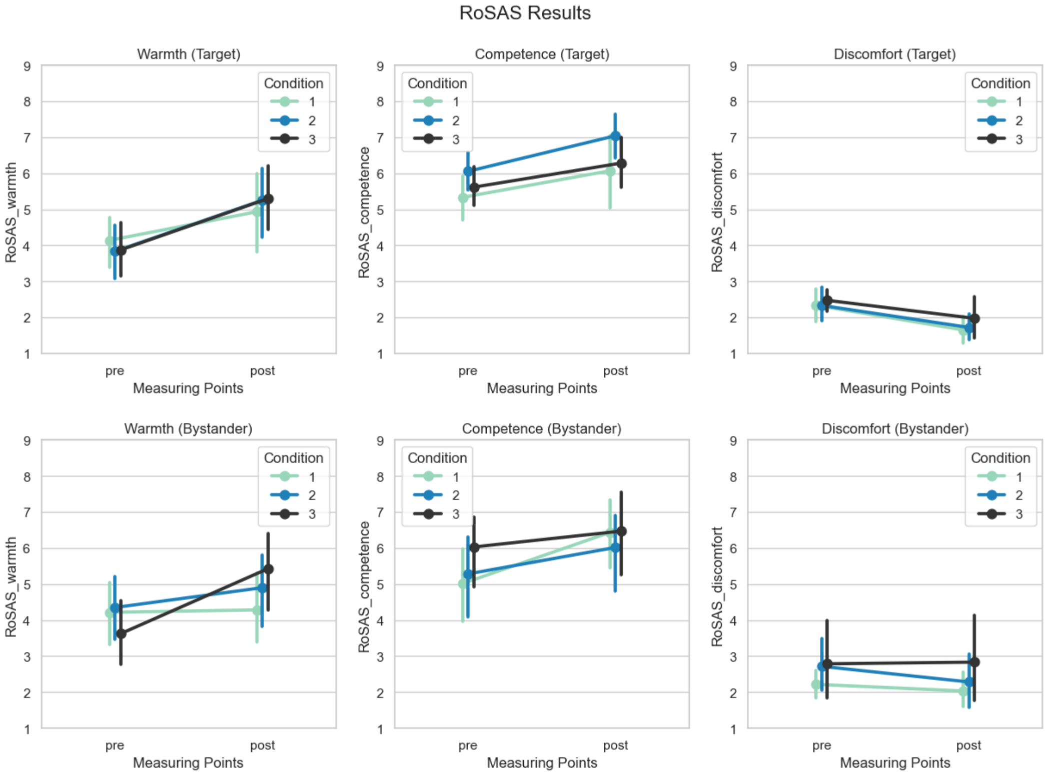


Figure 4.4: RoSAS Results Graph

Competence (*Target*). All assumptions for running an ANCOVA were met for the *target*. The covariate "Pepper error" was not significantly related to the pre-post assessments of competence of the *target* ($F(1,33) = 0.888, p > 0.05$, partial $\eta^2 = 0.026$). After controlling for the effect of the covariate "Pepper error", there was a significant increase in post measures of competence compared to pre ($F(1,33) = 14.168, p < 0.001$, partial $\eta^2 = 0.3$). There was no significant interaction effect ($F(2,33) = 0.239, p > 0.05$, partial $\eta^2 = 0.014$) and no significant between-subjects effect ($F(2,33) = 1.804, p > 0.05$, partial $\eta^2 = 0.099$).

Competence (*Bystander*). Normality was violated for the *bystander* position in the competence subscale. However, following the previous line of argumentation, I continued with the calculation as ANCOVAs are relatively robust against violations of normality (Vasey and Thayer, 1987). The covariate was not significantly related to the pre-post assessments of competence of the *bystander* ($F(1,27) = 0.257, p > 0.05$, partial $\eta^2 = 0.009$). After controlling for the covariate "Pepper error", there was a significant increase in the competence assessments ($F(1,27) = 8.076, p < 0.05$, partial $\eta^2 = 0.23$). There was no significant interaction effect ($F(2,27) = 1.398, p > 0.05$, partial $\eta^2 = 0.094$) and no between-subjects effect ($F(2,27) = 0.398, p > 0.05$, partial $\eta^2 = 0.029$).

Discomfort (*Target*). Normality was not given in the RoSAS discomfort scale. However, as argued before, due to the robustness of an ANCOVA (Vasey and Thayer, 1987), I continue with a mixed-model ANCOVA with the covariate "Pepper error" just as for the other RoSAS subscales. The covariate did not relate to the pre-post assessments of discomfort ($F(1, 33) = 0.459, p > 0.05$, partial $\eta^2 = 0.014$). After controlling for the effect of the covariate, there was a significant decrease in discomfort values ($F(1, 33) = 20.803, p < 0.001$, partial $\eta^2 = 0.387$). There was no significant interaction effect ($F(2, 33) = 0.077, p > 0.05$, partial $\eta^2 = 0.005$) and no significant between-subjects effect ($F(2, 33) = 0.612, p > 0.05$, partial $\eta^2 = 0.036$).

Discomfort (*Bystander*). As above, normality was not given. However, for the same argument, I continue with the ANCOVA. "Pepper error" did not relate to the pre-post assessment of discomfort ($F(1, 27) = 0.302, p > 0.05$, partial $\eta^2 = 0.011$). Controlling for the covariate, there was no significant difference between the pre-and post-assessments of discomfort for the *bystander* ($F(1, 27) = 0.706, p > 0.05$, partial $\eta^2 = 0.025$). There was no significant interaction effect ($F(2, 27) = 0.887, p > 0.05$, partial $\eta^2 = 0.062$) and no significant between subjects effect ($F(2, 27) = 0.798, p > 0.05$, partial $\eta^2 = 0.056$). Find an overview of all results of the RoSAS scale in Table 4.6.

Table 4.6: Results of the RoSAS scale with the three sub scales "Warmth", "Competence", "Discomfort".

		Within (team members)			**Interaction**			**Between (Cond)**		
	VP	*F*	*p*	η^2	*F*	*p*	η^2	*F*	*p*	η^2
War	1	31.107	<0.001*	0.485	1.465	0.246	0.082	0.010	0.990	0.001
	3	8.163	0.008*	0.232	3.277	0.053	0.195	0.116	0.890	0.009
Co	1	14.168	<0.001*	0.300	0.239	0.789	0.014	1.804	0.181	0.099
	3	8.076	0.008*	0.230	1.398	0.265	0.094	0.398	0.676	0.029
Dis	1	20.803	<0.001*	0.387	0.077	0.926	0.005	0.612	0.548	0.036
	3	0.706	0.408	0.025	0.887	0.424	0.062	0.798	0.460	0.056

War = Warmth, Co = Competence, Dis = Discomfort. VP1 stands for the *target* position, VP3 for the *bystander*. Significant values are highlighted with an *.

4.7 Conflict Perception

There were two subscales for conflict perception: 1. Relationship conflict, and 2. Task conflict. These scales were only collected after the intervention, so I initially planned to use One-Way ANCOVAs with the covariate "intentional sexism".

Relationship Conflict (*Target*). The homogeneity of regression slopes of the covariate "intentional sexism" for the *bystander* position was violated. Therefore, I

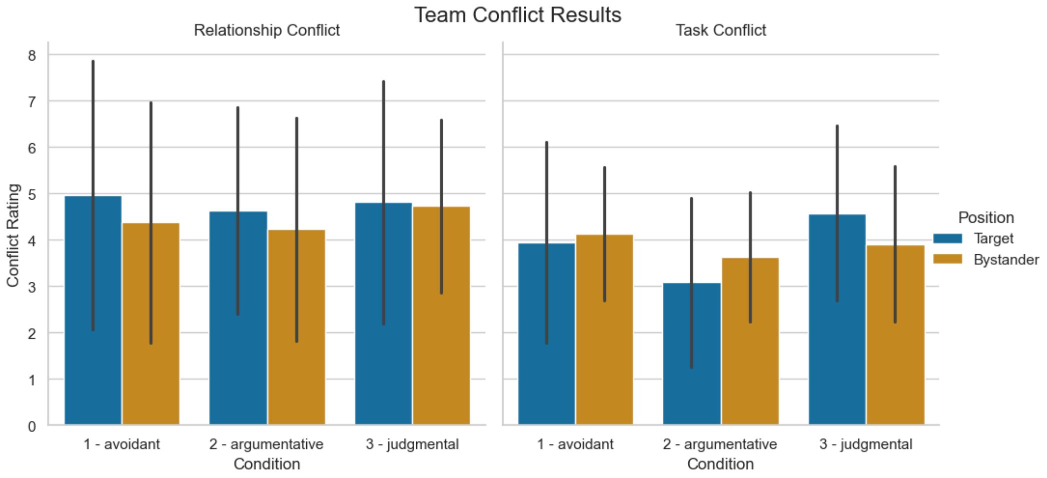


Figure 4.5: Results of Relationship Conflict and Team Conflict Scales

reduced the calculation to a One-Way ANOVA for both positions for the purpose of comparability. The results showed no significant difference between the conditions ($F(2,37) = 0.05, p > 0.05$, partial $\eta^2 = 0.003$) with $M = 4.96$ and $std = 2.89$ for condition 1, $M = 4.62$ and $std = 2.22$ for condition 2, and $M = 4.81$ and $std = 2.52$ for condition 3.

Relationship Conflict (*Bystander*). Calculating the ANOVA for relationship conflict for *bystander* also yielded no significant differences between the conditions ($F(2,31) = 0.134, p > 0.05$, partial $\eta^2 = 0.009$) with $M = 4.38$ and $std = 2.59$ for condition 1, $M = 4.23$ and $std = 2.40$ for condition 2, and $M = 4.73$ and $std = 1.87$ for condition 3.

Task Conflict (*Target*). In order to stay comparable between the two subscales, I decided to also run an ANOVA with task conflict. For *target*, normality was violated. However, since ANOVAs are robust against normality violations (Vasey and Thayer, 1987), I decided to proceed with the ANOVA. There were no significant differences between the different conditions for the *target* ($F(2,37) = 1.813, p > 0.05$, partial $\eta^2 = 0.096$) with $M = 3.94$ and $std = 2.16$ for condition 1, $M = 3.08$ and $std = 1.82$ for condition 2, and $M = 4.58$ and $std = 1.89$ for condition 3.

Task Conflict (*Bystander*). All assumptions were met for the *bystander*. The ANOVA found no significant differences between the three conditions ($F(2,31) = 0.273, p > 0.05$ partial $\eta^2 = 0.019$) with $M = 4.13$ and $std = 1.43$ for condition 1, $M = 3.62$ and $std = 1.40$ for condition 2, and $M = 3.91$ and $std = 1.69$ for condition 3.

Comparison of Conflict Scales (*Target*). I ran an additional ANOVA to see whether there were differences between the two subscales of task conflict and relationship conflict based on the different conditions. I found a significant difference between the two subscales for the *target* ($F(1,34) = 12.970, p < 0.05$, partial $\eta^2 = 0.276$) but no interaction effect ($F(2,34) = 2.213, p > 0.05$, partial $\eta^2 = 0.115$) or effect between the conditions ($F(2, 34) = 0.492, p > 0.05$, partial $\eta^2 = 0.028$).

Comparison of Conflict Scales (*Bystander*). For the bystander, the difference between the conflict scales was (barely) not significant ($F(1, 28) = 3.757, p > 0.05$, partial $\eta^2 = 0.118$). There was also no interaction effect ($F(2, 28) = 0.336, p > 0.05$, partial $\eta^2 = 0.023$) and no significant difference between the three conditions ($F(2, 28) = 0.144, p > 0.05$, partial $\eta^2 = 0.010$).

4.8 Closeness to other Team Members

I assessed the "Closeness to other Team Members" scale under consideration of the two covariates "confederate" and "intentional sexism". Normality assumptions were violated for both *target* and *bystander*. However, as argued before, ANCOVAs are relatively robust against these violations (Vasey and Thayer, 1987), so I proceeded with the mixed-model ANCOVA. This was not a pre-post assessment of one scale. Instead, I compared the same scale among the assessment of all three team members (the confederate, Pepper and the other participant, respectively) as the within-subjects component. The between-subjects components were the three conditions (i.e. avoidant, argumentative, and morally judgmental).

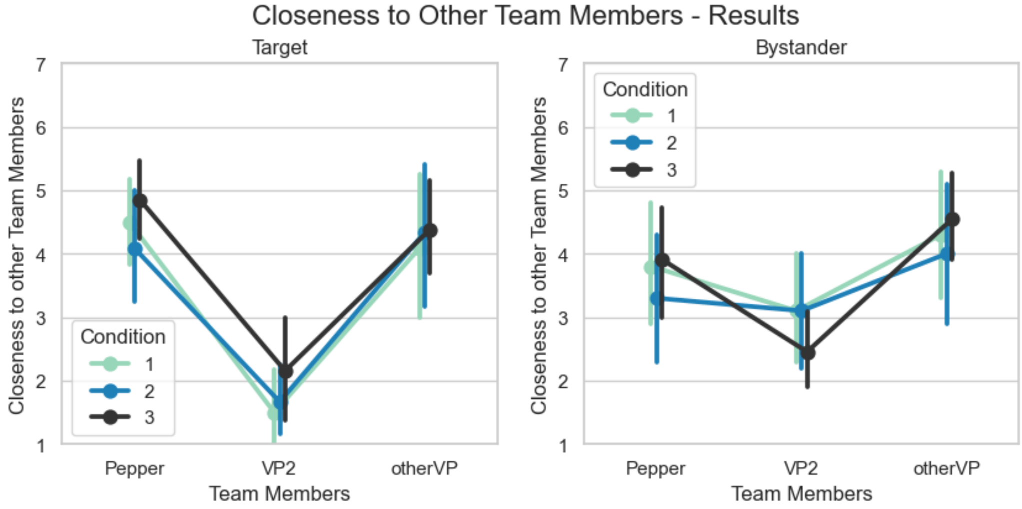


Figure 4.6: Results of Team Member Closeness

Closeness to Other Team Members (*Target*). The two covariates were not significantly related to the different people (within) that were judged (confederate: $F(2, 64) = 0.27, p > 0.05$, partial $\eta^2 = 0.008$; intentional sexism: $F(2, 64) = 1.112, p > 0.05$, partial $\eta^2 = 0.034$). After controlling for the effect of the covariates, there was a significant difference between the three team members being judged ($F(2, 64) = 4.884, p < 0.05$, partial $\eta^2 = 0.132$). Pairwise comparisons were conducted to examine differences in the relationship assessment between the other team members (within-subjects factor). The pairwise comparisons revealed a significant difference between Pepper (M(averaged over all three conditions)$= 4.486, std = 1.3$) and the sexist confederate ($M = 1.784, std = 1.205, p < 0.001$) as well as the sexist confederate and the other team member, the *bystander* ($M = 4.27, std = 1.82, p < 0.001$) in that the sexist confederate was evaluated significantly lower than the other two group members. There was no significant difference between the three conditions (between-subjects factor) ($F(2, 32) = 0.69, p > 0.05$, partial $\eta^2 = 0.041$) nor a significant interaction effect between the group member assessment and the conditions ($F(4, 64) = 0.439, p > 0.05$, partial $\eta^2 = 0.027$). See Table 8 in the appendix for all means and standard deviations.

Closeness to Other Team Members (*Bystander*). The two covariates were not significantly related to the different team members that were judged (confederate: $F(2, 52) = 1.429, p > 0.05$, partial $\eta^2 = 0.052$; intentional sexism: $F(2, 52) = 0.146, p > 0.05$, partial $\eta^2 = 0.006$). After controlling for the effect of the covariates, there was no significant difference between the three team members being evaluated ($F(2, 52) = 0.16, p > 0.05$, partial $\eta^2 = 0.006$). There was no significant difference between the conditions ($F(2, 26) = 0.158, p > 0.05$, partial $\eta^2 = 0.012$) nor a significant interaction effect ($F(4, 52) = 0.609, p > 0.05$, partial $\eta^2 = 0.045$).

Table 4.7: Results of Team Member Closeness

	Within (Team Members)			Interaction			Between (Cond)		
	F	p	η^2	F	p	η^2	F	p	η^2
VP1	4.884	0.011*	0.132	0.439	0.78	0.027	0.69	0.509	0.041
VP3	0.160	0.852	0.006	0.609	0.658	0.045	0.158	0.855	0.012

"VP1" stands for the *target*, "VP3" for the *bystander.*

4.9 Team Member Perception

As the "Team Member Perception" scale was a relatively new scale developed by Plum (2022), I first checked the reliability using Cronbach's Alpha (Cronbach, 1951) for all subscales. The reliability of the subscales "Sharing Mental Models" and "Viewing Interdependency as Positive" fell under the recommended threshold of 0.7 (Tavakol and Dennick, 2011). Therefore, I excluded those subscales from the calculation. The reliability for the "Knowing and Fulfilling their Roles" subscale for the *target*'s assessment of the confederate also was bad ($\alpha = 0.591$). However, as all other reliability assessments of that scale were reliable, I decided to include it, interpreting it with caution. All other subscale's reliabilities were sufficient. Find an exhaustive list of all reliability values in Table 4.8.

I calculated mixed-model ANCOVAs for all subscales of the Team Member Perception scale. I used the covariates "confederate" and "intentional sexism". Normality

Table 4.8: Reliability measured with Cronbach's Alpha of the Team Member Perception

		Cronbach's Alpha		
		Pepper	**VP2**	**otherVP**
Sharing Mental Models	*Target*	-0.063*	0.550*	-0.473*
	Bystander	0.845	0.427*	0.467*
Subjugating Individual Needs for Group Needs	*Target*	0.739	0.685	0.749
	Bystander	0.906	0.700	0.805
Viewing Interdependency as Positive	*Target*	0.290*	0.730	0.225*
	Bystander	0.596*	0.795	0.580*
Knowing and Fulfilling their Roles	*Target*	0.731	0.591*	0.742
	Bystander	0.858	0.797	0.897
Trust	*Target*	0.816	0.845	0.781
	Bystander	0.917	0.896	0.918
Social Interaction	*Target*	0.775	0.898	0.738
	Bystander	0.901	0.869	0.874

"VP2" refers to the confederate. All scales with insufficient reliability are marked with an asterix (*).

was violated for all subscales. However, for the same reasons as stated before (Vasey and Thayer, 1987), I continued with the ANCOVA's calculation.

Subjugating Individual Needs for Group Needs (*Target*). Box's M test of the equality of covariance matrices (Box, 1949) was significant ($p < 0.05$). Therefore, I excluded the covariates from the calculation and performed an ANOVA. (Contrary to the previous decision to continue with the simpler analysis method for all following calculations of the scale, I this time decided to continue with the ANCOVA for all other subscales, as no other subscale violated the assumptions, and there were multiple subscales where it was interesting to see the effect of the covariates). Mauchly's test of sphericity (Mauchly, 1940) was significant. Therefore, I used Greenhouse-Geisser corrections (Geisser and Greenhouse, 1958). There was a significant difference between the assessments of the three different team members ($F(2, 68) = 26.446, p < 0.001$, partial $\eta^2 = 0.438$). Pairwise comparisons resulted in a significant difference between the sexist confederate ($M = 4.14, std = 1.03$) and Pepper ($M = 5.21, std = 0.92, p < 0.001$), as well as the sexist confederate and the *target* ($M = 5.033, std = 0.96, p < 0.001$). There was no significant interaction effect ($F(4, 68) = 1.397, p > 0.05$, partial $\eta^2 = 0,076$). However, there was a significant main effect in the difference between the three conditions ($F(2, 34) = 3.452, p < 0.05$, partial $\eta^2 = 0.169$). Tukey's post-hoc test (Tukey, 1949) showed a significant difference between condition 2 (argumentative)($M = 5.25$) and 3 (morally judgmental)($M = 4.5, p < 0.05$), in that condition 2 was rated significantly higher than condition 3 across all three different team members that were assessed. See all means in the appendix in Table 9. Figure 4.7 provides an overview of the means for the first two subscales of the Teammember Perception Scale.

Subjugating Individual Needs for Group Needs (*Bystander*). All assumptions were met (apart from normality). There was no significant relationship between the covariates and the assessments of the different team members (confederate: $F(2, 52) = 0.102, p > 0.05$, partial $\eta^2 = 0.04$; intentional sexism: $F(2, 52) = 0.511, p > 0.05$, partial $\eta^2 = 0.019$). After controlling for the covariates, there was no significant difference between the assessments of the three different team members ($F(2, 52) = 0.197, p > 0.05$, partial $\eta^2 = 0.008$). There was also no significant interaction effect ($F(4, 52) = 0.632, p > 0.05$, partial $\eta^2 = 0.046$). There was a significant relationship between the covariate "confederate" and the assessment of the different conditions ($F(1, 26) = 5.793, p < 0.05$, partial $\eta^2 = 0.182$). Controlling for the covariates, there was no significant difference between the different conditions ($F(2, 26) = 0.399, p > 0.05$, partial $\eta^2 = 0.03$).

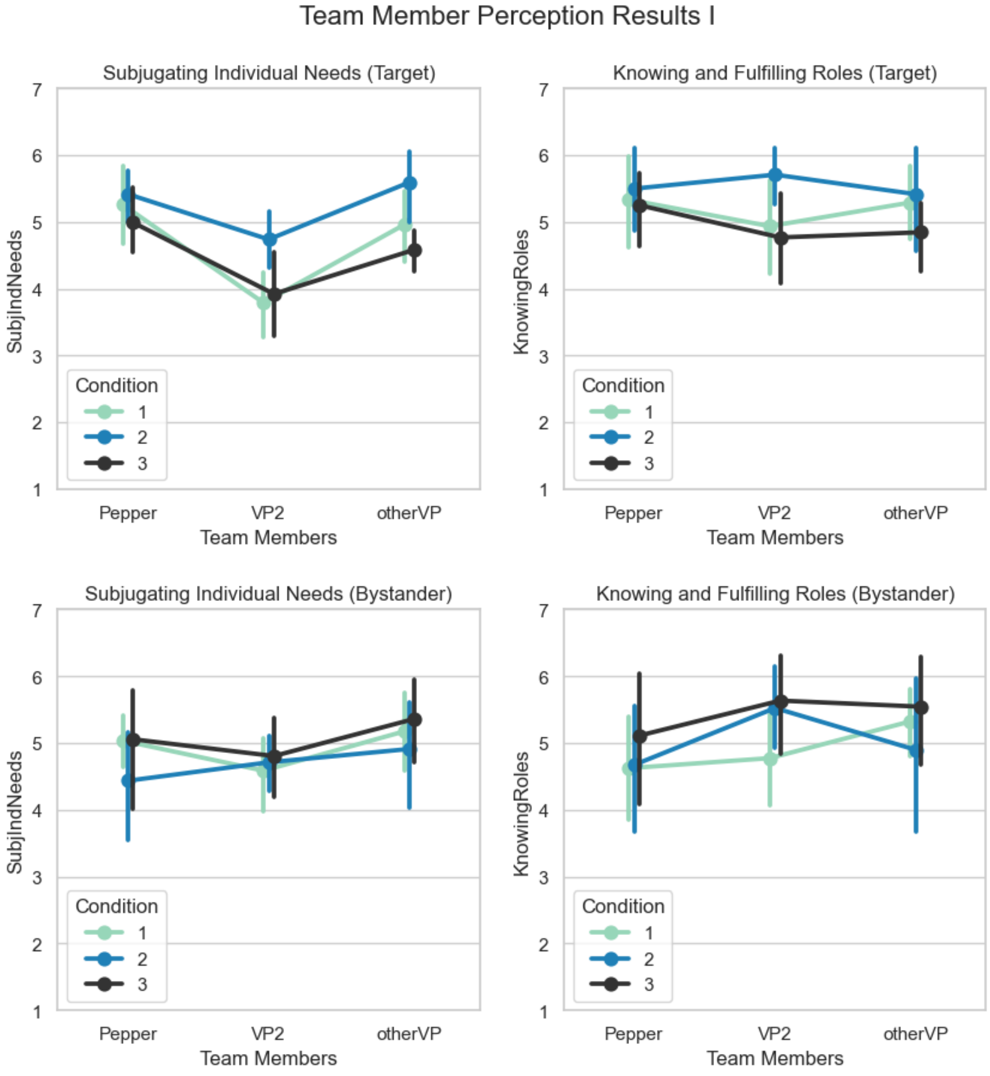


Figure 4.7: Graphic of the Results for Subscales "Subjugating Individual Needs for Group Needs" and "Knowing and Fulfilling their Roles"

Knowing and Fulfilling their Roles (*Target*). All assumptions (apart from normality) were met. Therefore, I calculated an ANCOVA. There was no significant relationship of the covariates with the dependent variable (confederate: $F(2, 64) = 0.064, p > 0.05$, partial $\eta^2 = 0.002$; intentional sexism: $F(2, 64) = 1.445, p > 0.05$, partial $\eta^2 = 0.043$). After controlling for the effect of the covariates, there was no significant difference between the assessment of the different team members ($F(2, 64) = 0.086, p > 0.05$, partial $\eta^2 = 0.003$). There was also no significant interaction effect between team members and the conditions ($F(4, 64) = 0.687, p > 0.05$, partial $\eta^2 = 0.041$), nor was there a significant difference between the three conditions ($F(2, 32) = 1.702, p > 0.05$, partial $\eta^2 = 0.096$).

Knowing and Fulfilling their Roles (*Bystander*). The assumption of homogeneity of regression slopes was violated for the covariate "intentional sexism". Therefore, I calculated an ANCOVA only with the covariate "confederate". Here, all assumptions were met. "Confederate" was not significantly related to the assessment of the different team members and conditions ($F(2, 54) = 0.479, p > 0.05$, partial $\eta^2 = 0.017$). After controlling for the covariate, there was no significant difference between the assessment of the three team members ($F(2, 54) = 0.260, p > 0.05$, partial $\eta^2 = 0.01$), nor was there an interaction effect ($F(4, 54) = 0.785, p > 0.05$, partial $\eta^2 = 0.055$) or a significant difference between the three conditions ($F(2, 27) = 0.602, p > 0.05$, partial $\eta^2 = 0.043$). See Figure 4.7 for a graphical overview of the first two subscales.

Trust (*Target*). Mauchly's test of sphericity (Mauchly, 1940) was significant ($p < 0.001$). Therefore, I used the Greenhouse-Geisser correction (Geisser and Greenhouse, 1958) for a corrected assessment of the within-subjects effect. None of the covariates were significantly related to the within-subjects measurements (confederate: $F(2, 64) = 0.092, p > 0.05$, partial $\eta^2 = 0.003$; intentional sexism: $F(2, 64) = 0.443, p > 0.05$, partial $\eta^2 = 0.014$). After controlling for the effect of the covariates, there was a significant difference between the three team members being assessed (Pepper, sexist confederate, *bystander*)($F(2, 64) = 4.149, p < 0.05$, partial $\eta^2 = 0.115$). Pairwise comparisons showed a significant difference between Pepper ($M = 4.86, std = 1.06$) and the sexist confederate ($M = 3.12, std = 1.39, p < 0.001$), as well as the sexist confederate and the *bystander* ($M = 4.69, std = 1, p < 0.001$). The covariate "confederate" was significantly related to the between-subjects values ($F(1, 32) = 5.459, p < 0.05$, partial $\eta^2 = 0.146$). After controlling for the covariates, there was no significant difference between the different conditions ($F(2, 32) = 0.945, p > 0.05$, partial $\eta^2 = 0.056$), nor a significant interaction effect between the team members and the different conditions ($F(4, 64) = 1.174, p > 0.05$, partial $\eta^2 = 0.068$).

Trust (*Bystander*). For the trust measurement of the *bystander* position, Levene's test of equality of error variances (Levene, 1960) was violated. However, following Field (2013), the ratio of variances was below the critical value of approximately 5 ($1.58^2/0.91^2 = 3.04$) that is listed for three groups of approximately ten people. Therefore, I continued with the calculation of the ANCOVA. There was no significant relationship between the covariates and the within-subjects values (confederate: $F(2, 52) = 0.230, p > 0.05$, partial $\eta^2 = 0.009$; intentional sexism:

$F(2, 52) = 0.247, p > 0.05$, partial $\eta^2 = 0.009$). After controlling for the covariates, there was no significant difference between the different team members (within-subjects component) ($F(2, 52) = 1.171, p > 0.05$, partial $\eta^2 = 0.043$), nor significant differences between conditions ($F(2, 26) = 1.452, p > 0.05$, partial $\eta^2 = 0.1$), nor was there a significant interaction effect ($F(4, 52) = 0.438, p > 0.05$, partial $\eta^2 = 0.033$). However, looking at pairwise comparisons, I found a significant difference between the confederate ($M = 3.66, std = 1.39$) and the other participant, i.e. the *target* ($M = 4.45, std = 1.38, p < 0.05$). See Figure 4.8 for a graphical overview of the results of the scale.

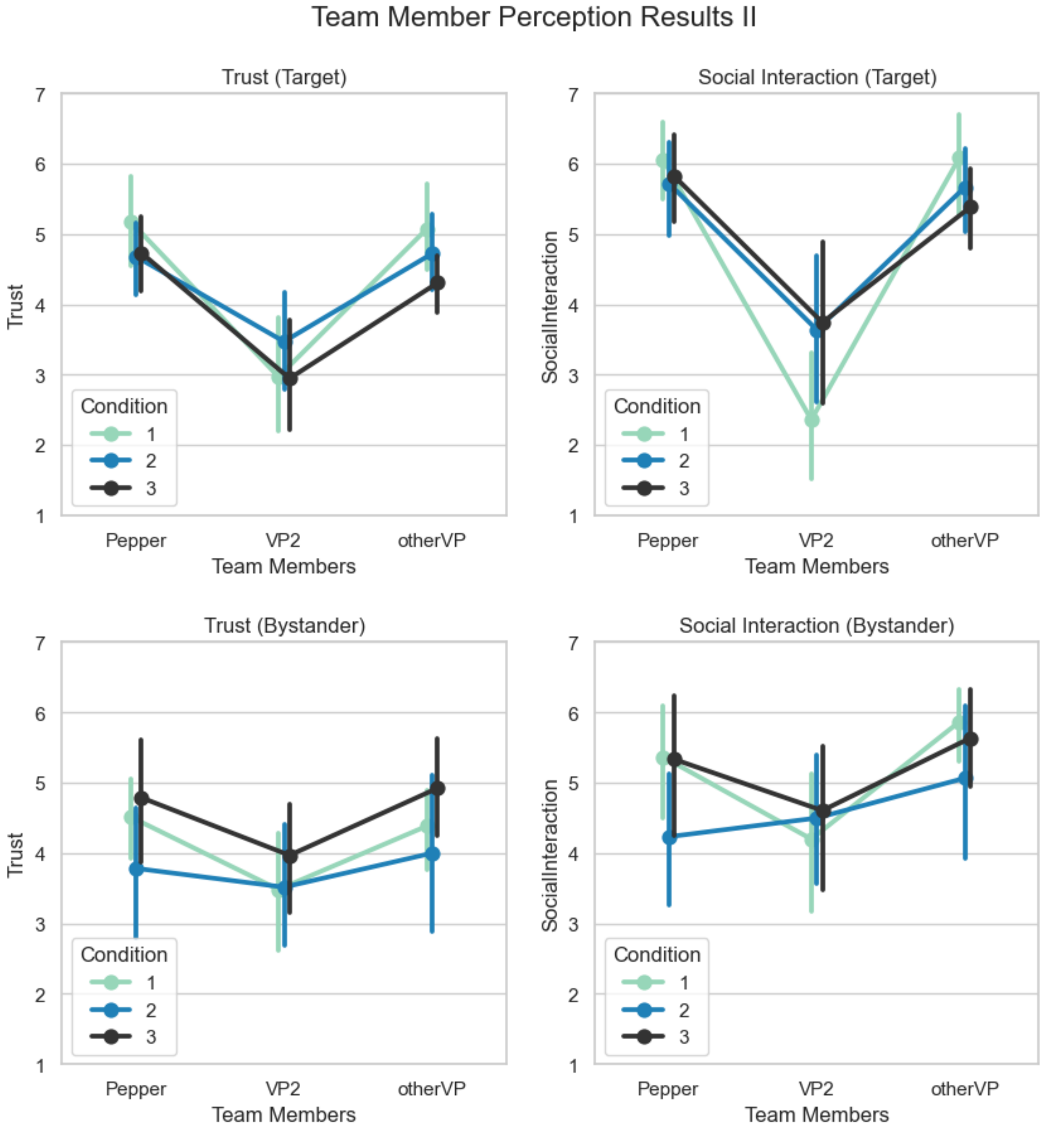


Figure 4.8: Graphic of the Results for Subscales "Trust" and "Social Interaction"

Social Interaction (*Target*). Mauchly's test of sphericity (Mauchly, 1940) was significant ($p < 0.05$), therefore I used the Greenhouse-Geisser correction (Geisser and Greenhouse, 1958). There was no significant relationship of the covariates with the assessment of the different team members (confederate: $F(2,62) = 0.092, p > 0.05$, partial $\eta^2 = 0.003$; intentional sexism: $F(2,64) = 2.69, p > 0.05$, partial $\eta^2 = 0.078$). Controlling for the covariates, I found a significant difference between the three team members ($F(2,64) = 4.682, p < 0.05$, partial $\eta^2 = 0.128$). Pairwise comparisons showed a significant difference between the sexist confederate ($M = 3.26, std = 2.0$) and Pepper ($M = 5.86, std = 1.12, p < 0.001$), as well as the sexist confederate and the *bystander* ($M = 5.7, std = 1.2, p < 0.01$). There was no significant difference between the three conditions ($F(2,32) = 0.144, p > 0.05$, partial $\eta^2 = 0.009$). However, there was a significant interaction effect ($F(4,64) = 3.505, p < 0.05$, partial $\eta^2 = 0.18$) where the assessment of the confederate in the avoidant condition seems significantly lower than the assessments of the confederate in the argumentative and morally judgmental condition. See Figure 4.8 for reference.

Social Interaction (*Bystander*). All assumptions were met for the *bystander* position (apart from normality as mentioned above). The covariate "confederate" was significantly related to the assessment of the different conditions ($F(1,26) = 9.323, p < 0.05$, partial $\eta^2 = 0.264$). Controlling for the effect of the two covariates, there was no significant difference between the assessment of the different team members (within-subjects component) ($F(2,52) = 0.002, p > 0.05$, partial $\eta^2 = 0$). There also was no significant difference between the conditions ($F(2,26) = 1.282, p > 0.05$, partial $\eta^2 = 0.09$), nor was there a significant interaction between team members and conditions ($F(4,62) = 0.981, p > 0.05$, partial $\eta^2 = 0.07$).

Table 4.9: Results of "Team Member Perception" Scale

		Within (team members)			Interaction			Between (Cond)		
		F	*p*	η^2	*F*	*p*	η^2	*F*	*p*	η^2
Subj	**VP1**	26.446	<0.001**	0.438	1.397	0.253	0.076	3.452	0.043*	0.169
Needs	**VP3**	0.197	0.822	0.008	0.632	0.642	0.046	0.399	0.675	0.030
Know.	**VP1**	0.086	0.918	0.030	0.687	0.604	0.041	1.702	0.198	0.096
Roles	**VP3**	0.260	0.772	0.010	0.785	0.540	0.055	0.602	0.555	0.043
Trust	**VP1**	4.149	0.033*	0.115	1.174	0.329	0.068	0.945	0.399	0.056
	VP3	1.171	0.318	0.043	0.438	0.780	0.033	1.452	0.252	0.100
Soc.	**VP1**	4.682	0.022*	0.128	3.505	0.022*	0.180	0.144	0.866	0.009
Inter.	**VP3**	0.002	0.998	0.000	0.981	0.426	0.070	1.282	0.294	0.090

"VP1" refers to the *target* position, "VP3" refers to the *bystander* position. η^2 refers to "partial η^2". These were shortened to fit on the page. * marks significant results below 0.05, ** results below 0.001 respectively.

4.10 Qualitative Analysis

For the qualitative analysis, both the video recordings as well as the interviews were analysed.

4.10.1 Video Analysis

Immediate Reaction after Pepper's Intervention. Most importantly, I analysed how people responded to Pepper's intervention. For this, I looked at the immediate reaction and deducted different responses following Mayring and Fenzl (2019). As some reactions were hard to discern from another, it was possible to code a maximum of two categories per person. In total, 110 reactions were considered: 31 for condition 1 (*target*: 15, *bystander*: 16), 41 for condition 2 (*target*: 23, *bystander*: 18) and 38 for condition 3 (*target*: 21, *bystander*: 17).
The most common reactions in the avoidant condition were to "continue the game" (7/31 = 23% in total: *target*: 3/15 = 20%; *bystander*: 4/16 = 25%) or "look at the other team members" (6/31 = 19% in total: *target*: 3/15 = 20%; *bystander*: 3/16 = 19%). The most common reactions in the argumentative condition were to laugh (10/41 = 24% in total: *target*: 5/23 = 22%; *bystander*: 5/18 = 28%) or also to look at the other team members (10/41 = 24% in total: *target*: 3/23 = 13%; *bystander*: 7/23 = 30%). The most common reactions for the morally judgmental condition were to laugh (9/38 = 24% in total: *target*: 3/21 = 14%; *bystander*: 6/17 = 35%) or to look at the confederate (8/38 = 21% in total: *target*: 5/21 = 24%, *bystander*: 3/17 = 18%). See Table 4.10 for a detailed listing of all responses.

Table 4.10: Overview of Immediate Reactions after Pepper's Intervention

	Avoidant			**Argumentative**			**Judgmental**			**Count**		
Reaction	VP1	VP3	Σ	VP1	VP3	Σ	VP1	VP3	Σ	VP1	VP3	Total
Laughing	1	3	4	5	5	10	3	6	9	9	14	23
Eyeing team	3	3	6	3	7	10	4	2	6	10	12	22
Eyeing conf.	3	1	4	6	1	7	5	3	8	14	5	19
Pepper inter.	2	2	4	6	1	7	4	2	6	12	5	17
Continue game	3	4	7	2	1	3	3	1	4	8	6	14
No reaction	2	2	4	1	0	1	1	1	2	4	3	7
Inter. conf.	1	0	1	0	3	3	1	2	3	2	5	7
Total	15	16	31	23	18	41	21	17	38	59	51	110

The abbreviation "conf." stands for "confederate". The abbreviation "inter." stands for "interaction". "VP1" refers to the *target*, and "VP3" to the *bystander*. These were shortened for space purposes.

Initiating the Comment. Other information I gathered from the video interviews was whether the *target* had proposed the comment, upon which she would receive the sexist comment, as this might influence the participant's reactions and perceptions as well. Most of the targets initiated their comment (19/36 = 53%), the confederate asked ten *target* people (28%) to give their thoughts, and sometimes the robot Pepper had to encourage the *target* person to say something, as all the other options did not work to get the *target* to speak as a basis for the sexist comment (7/36 = 19%).

Treatment of Pepper. Regarding people's initial treatment of Pepper, only a few addressed Pepper directly, for example, by asking questions (11/68 = 16%). Whereas most reacted non-verbally to Pepper's suggestions (50/68 = 74%), some reacted to Pepper's suggestions by answering verbally directed at Pepper (7/68 = 10%).

4.10.2 Interview Analysis

Emotional Responses. Responses to the study were quite strong. N = 3 (8%) of the *target* people started crying in the interview or during the debriefing when the tension subsided. Generally, people most often reported feeling angry (16/68 = 24%) or irritated (20/68 = 29%). This was similar for condition 1 (angry: 7/22 = 31%; irritated: 6/22 = 27%) and condition 2 (angry: 6/22 = 27%; irritated: 6/22 = 27%). In condition 3, this looked a little different (angry: 3/24 = 13%; irritated: 8/24 = 33%). There were also differences between the *target* (angry: 12/37 = 32%; irritated: 10/37 = 27%) and the *bystander* (angry: 4/31 = 13%; irritated: 10/31 = 32%). Taken together, being angry or irritated across the conditions decreased with each condition for the *target* (condition 1: 10/12 = 83%; condition 2: 7/12 = 58%; condition 3: 5/13 = 38%) whereas those emotions increased with each condition for the *bystander* (condition 1: 3/10 = 30%, condition 2: 5/10 = 50%, condition 3: 6/11 = 55%).
Other findings show that *bystanders* were more surprised and in shock about the situation (8/31 = 26%) than *target* people (4/37 = 11%). Some only expressed a minor annoyance (*target*: 3/37 = 8%; *bystander*: 2/31 = 6%) or did not care (*target*: 3/37 = 8%; *bystander*: 5/31 = 16%), often justified with frequent similar experienced situations in their past. Some had not heard the comment and, therefore, did not express many emotions or instead talked positively about the experience (*target*: 4/37 = 11%; *bystander*: 1/31 = 3%). One person from each position thought the whole intervention was primarily funny.

Mentioning Sexist Encounter. I assessed whether people immediately mentioned the sexist encounter when asked about their experience in the experiment. Mostly, they did immediately mention the sexist encounter, one of them even before the actual start of the interview: Condition 1: 16/22 = 73% (*target*: 9/12 = 75%, *bystander*: 7/10 = 70%), condition 2: 18/22 = 82% (*target*: 10/12 = 83%, *bystander*: 8/10 = 80%), condition 3: 11/24 = 46% (*target*: 4/13 = 31%, *bystander*: 7/11 = 64%). However, for the morally judgmental condition (3rd condition), many people had to explicitly be asked about the encounter to tell of it (total: 8/24 = 33%; *target*: 5/13 = 38%, *bystander*: 3/11 = 27%), whereas in the other conditions, this was 4/22 (= 18%) for the avoidant condition, and 3/22 (= 14%) for the argumentative, all evenly distributed across the *bystander* and *target* position. Nine participants reported being taken entirely by surprise and never having experienced anything like that. All other people did not mention the sexist encounter (condition 1: 2/22 = 9%; condition 2: 1/22 = 5%; condition 3: 4/24 = 17%).

Interpreting Comment as Sexism. Seven people in each position (*target/bystander*) mentioned that they were wondering whether the sexist comment was "real". This was distributed unequally across the conditions, as it was only two in the avoidant condition (9%), equally across positions, four (18%) in the argumentative condition (*target*: 3/12 = 25%, *bystander*: 1/10 = 10%) and eight (33%) in the morally judgmental condition (*target*: 3/13 = 23%, *bystander*: 5/11 = 45%). One person in each position (*target*/*bystander*) pondered whether the confederate had voiced the sexist comment on purpose to test how Pepper would react to this. I included the information on people interpreting the comment in the intended way or not as a covariate in the quantitative statistical analysis in chapter 4.

Experience of Pepper's Intervention. The next question was regarding how the participants had experienced Pepper's intervention. More than half of all participants viewed Pepper's intervention as helpful (35/68 = 51%). However, there was a clear distinction between the conditions. Only three *target* people of the avoidant condition (25%) viewed Pepper's intervention as helpful, in contrast to 17/22 = 77% (*target*: 11/12 = 92%, *bystander*: 6/10 = 60%) in the argumentative condition and 15/24 = 63% (*target*: 8/13 = 62%, *bystander*: 7/11 = 64%) in the morally judgmental condition. Instead, most people in the avoidant condition did not recall noticing that Pepper had intervened in any way (16/22 = 73%, *target*: 8/12 = 67%, *bystander*: 8/10 = 80%). In contrast, only two of the argumentative condition (9%), both *bystanders*, did not mention Pepper intervening, and five of the morally judgmental condition (*target*: 4/13 = 31%, *bystander*: 1/11 = 9%). After debriefing the people from the avoidant condition about Pepper's intervention, some of them mentioned that they had thought that Pepper's statement referred to the game and that the robot did not agree with what one of the team members had proposed. However, as these comments were only mentioned after the study run, I cannot provide a concrete number here, but I would like to add it as anecdotal evidence.

Reflection on Own Behaviour. Only 35 (51%) provided reasons for their behaviour. Most common responses were that they thought intervening would not be helpful (9/35 = 26% of the people providing reasons; target: 6/20 = 30%; bystander: 3/15 = 20%) or that they were too overwhelmed to react (9/35 = 26% of the people providing reasons; target: 4/20 = 20%; bystander: 5/15 = 33%). For being overwhelmed, this was almost evenly distributed across conditions. For the response that it would not be helpful there were slight differences between the conditions (condition 1: target: 2/6 = 33%; bystander: 1/4 = 25%; condition 2: target: 0%; bystander: 1/4 = 25%; condition 3: target: 4/9 = 44%; bystander: 1/7 = 14%) with the target position being more discouraged in the avoidant and morally judgmental condition. Three people said they wished they had intervened. Two people said they did not know the person and therefore could not grasp whether it was a joke. Another two people said there was limited time, and they had to solve the task. One person said she wanted to discuss this later with the confederate. Another person said it would have been uncomfortable to address the issue. One bystander of the morally judgmental condition said that Pepper had already intervened, so it was not necessary to add anything anymore.

Wishes for Future Incidents. Due to the semi-structured nature of the interview, not all participants answered all questions, which was also dependent on their affectedness of the incident. Regarding wishes on how to handle such a situation better in the future, I had 26 people (38% in total; target: 10/37 = 27%; bystander: 16/31 = 52%) that did not answer this question. However, those that answered this most often wished for other people to intervene, some even referring to how Pepper had intervened (condition 1: target: 4/12 = 33%; bystander: 4/10 = 40%; condition

2: target: $5/12 = 42\%$; bystander: $2/10 = 20\%$; condition 3: target: $6/13 = 47\%$; bystander: $5/10 = 50\%$). One person also specifically mentioned that making eye contact with the bystander position had supported her. Wishes regarding reacting how Pepper did were more common in the argumentative and morally judgmental condition than in the avoidant condition (condition 1: $2/22 = 9\%$; condition 2: $4/22 = 18\%$; condition 3: $5/24 = 21\%$). One male bystander even wrote down the different answers of Pepper as a reference for future incidents. When explicitly asked whether they would have wished for a response from Pepper, seven people in the avoidant condition ($7/22 = 32\%$) said they could not imagine or did not expect Pepper to intervene on an interpersonal level. $N = 5$ (14%) of the target people wished for more friendly team communication to avoid any such incident in the first place. A male bystander openly reflected that he now knew how he would like to answer instead. Some participants were thankful for the good opportunity to learn. As those comments fell after the debriefing upon the realisation that it was all part of the study, I cannot provide precise numbers on this either, as I did not note all the comments that participants mentioned then.

Pepper's Implicit Gender. I assessed which pronoun participants used to describe Pepper to imply which gender they assigned Pepper subconsciously. Participants most often used the male pronoun to describe Pepper ($35/68 = 51\%$ in total: *target*: $18/37 = 49\%$; *bystander*: $17/31 = 55\%$), followed by the female pronoun ($17/68 = 25\%$ in total: *target*: $11/37 = 30\%$; *bystander*: $6/31 = 19\%$). The rest, either intentionally or unintentionally, did not mention any pronouns, or raised their concern of assigning Pepper a gender. In the first two conditions the male pronoun was the most prominent (condition 1: $15/22 = 68\%$ in total *target*: $9/12 = 75\%$; *bystander*: $6/10 = 60\%$; condition 2: $14/22 = 64\%$ *target*: $6/12 = 50\%$; *bystander*: $8/10 = 80\%$). In contrast the female pronoun: Condition 1: $4/22 = 18\%$ in total *target*: $3/12 = 25\%$; *bystander*: $1/10 = 10\%$; Condition 2: $5/22 = 23\%$ *target*: $3/12 = 25\%$; *bystander*: $2/10 = 20\%$. In the morally judgmental condition, the gender attribution was more balanced between male, female and no pronoun used (female: $8/24 = 33\%$ in total, *target*: $5/13 = 38\%$, *bystander*: $3/11 = 27\%$). The male pronoun was used six times, as was "no pronoun" (25%), both equally often for *target* and *bystander*.

Pepper's Explicit Gender. When being asked explicitly about Pepper's gender, most people said that Pepper had no gender ($27/68 = 40\%$ in total: *target*: $15/37 = 41\%$, *bystander*: $12/31 = 39\%$, closely followed by Pepper being viewed as male ($26/68 = 38\%$ in total: *target*: $12/37 = 32\%$, *bystander*: $14/31 = 45\%$). Half as many people said Pepper was female ($15/68 = 22\%$ in total; *target*: $10/37 = 27\%$, *bystander*: $5/31 = 16\%$). Looking at the different conditions, in condition 1, most people viewed Pepper as male ($13/22 = 60\%$ in total: *target*: $6/12 = 50\%$, *bystander*: $7/10 = 70\%$), followed by neutral ($7/22 = 32\%$ in total: *target*: $5/12 = 41\%$, *bystander*: $2/10 = 20\%$). Only two people (9%) viewed Pepper as female, evenly across positions. For condition 2, most people viewed Pepper as neutral ($9/22 = 41\%$ in total: *target*: $3/12 = 25\%$, *bystander*: $6/10 = 60\%$) or male ($8/22 = 36\%$ in total; *target*: $5/12 = 42\%$, *bystander*: $3/10 = 30\%$). The remaining five people (23%) viewed Pepper as female; four of them were in the *target* position. In condition 3,

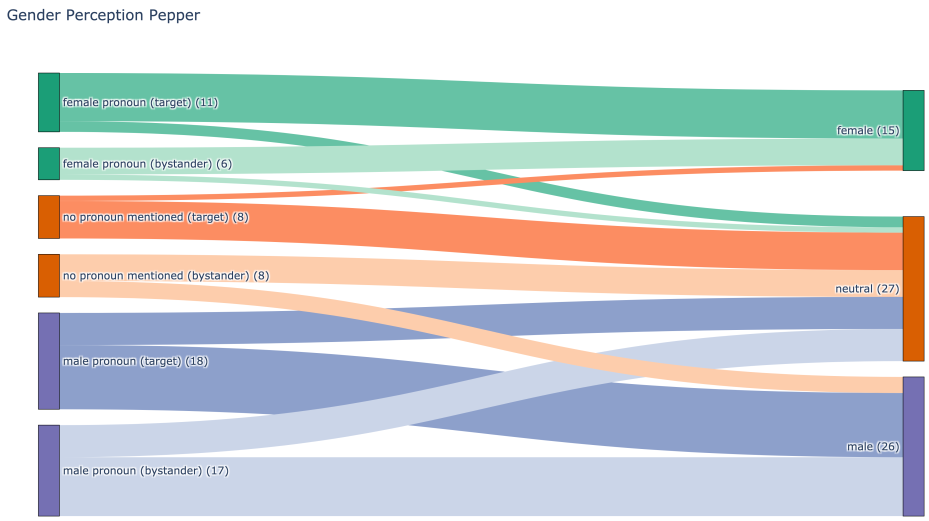


Figure 4.9: Sankey diagram of people's usage of pronouns when describing Pepper and people's perceived gender of Pepper.

most people viewed Pepper as neutral ($11/24 = 46\%$ in total; *target*: $7/13 = 54\%$, *bystander*: $4/11 = 36\%$), followed by female ($8/24 = 32\%$ in total; *target*: $5/13 = 38\%$, *bystander*: $3/13 = 23\%$). See Figure 4.9 for a depiction of how implicit use of pronouns was connected with explicit attribution of gender.

Reasoning for Pepper's Gender. Most participants (85%) explained why they assigned Pepper a certain gender. Most common ($17/58 = 29\%$) was the response that robots simply were neutral objects. Eight people ($8/58 = 14\%$) thought Pepper was a male name. Seven people ($7/58 = 12\%$) said Pepper was a female name. One of them referred to a Marvel movie in which there was a female assistant called Pepper. Further seven people ($7/58 = 12\%$) said they inferred the gender from the German article for "The Robot" (="der Roboter"). Five people ($5/58 = 9\%$) viewed the voice as female, and four ($4/58 = 7\%$) as male. Further six people ($6/58 = 10\%$) said the figure of the robot seemed male or that the fact that Pepper had no hair made them assess Pepper as male. Two people ($2/58 = 3\%$) mentioned that they believed their gender assessment of the robot might have been influenced by the situation, either believing Pepper to be female as the robot was regarded as an ally against sexism. Alternatively, as the group composition was two male versus one female member, they believed that Pepper being female would equalise this.

Chapter 5

Discussion

This study aimed to investigate which impact a robot could have when intervening in sexist encounters in human-robot group settings. For this, three different responses were tested: avoidant, argumentative and morally judgmental. I assessed whether these responses had an impact on positive and negative affect, the social attributions of the robot, team conflict perception, as well as perceptions of how the different team members contributed to the team and fulfilled specific criteria, making up good team members. I conducted a qualitative interview to better understand the behaviour witnessed in the interaction. I analysed the effects for the person being the *target* of the comment and the person just witnessing the event, the *bystander*.

5.1 Discussion of Results

5.1.1 Direct Reaction after Interaction

To answer RQ2, whether repair comments by a robot empowered people affected by sexism, as well as RQ3, whether there were differences between being the target or the bystander, I analysed the 49 video recordings of the experiment. Here, I found interesting differences between the conditions. Whereas the most common response with 24% in the argumentative and morally judgmental condition was laughter, this percentage was halved in the avoidant condition. The most common response here was to directly continue the game or stare at the other team members, potentially in search of help or purely out of shock. In the morally judgmental and argumentative condition, many other people also accusingly stared at the confederate or directly interacted with Pepper in response. Here again, this was around half as common in the avoidant condition. Combining this with findings from the interviews that in the avoidant condition, 73% did not notice that Pepper had said something about the incident, lets me conclude that people might have felt left alone in this encounter and decided to ignore the sexist comment to not further escalate the conflict, both in the target and bystander position. They also did not speak up as much or interacted with Pepper, nor did they stare at the sexist confederate. All of these behaviours might have required more strength and might also have involved more risk. This is in line with findings by Swim and Hyers (1999), who stress the difficulty of speaking up in these situations. Many participants reported in the interview that they were taken completely by surprise and had never experienced anything like that. They, however, often wished that the bystander person would have said something. In case the bystander did, they often appreciated this. People in the argumentative and morally judgmental conditions often referred to Pepper's response and said that this is what they would wish for as a response.

Therefore, based on the qualitative data, people might not recognise the avoidant condition as an intervention. Perhaps they do not expect Pepper to react to interpersonal conflicts and, therefore, do not attribute Pepper's reaction to be such an intervention. This goes in line with participants stating that they thought Pepper's comment was directed at the proposal of one team member regarding how to continue the game. This, however, goes against findings by Edwards et al. (2019) that people expect to interact with humans when interacting with social robots. One could argue that if they expected Pepper to behave human-like, more participants would have recognised the hint regarding the interpersonal conflict. Or, potentially, as Edwards et al. find, participants adapted their expectations rather quickly upon initial contact and, as a result, did not expect the robot to react this way. So, if the goal is to empower women, it seems that an avoidant response as endorsed as the minimal response by West et al. (2019) does not suffice in countering sexism – at least when measured by how many people notice the interaction and how actively they respond. Instead, addressing the norm violation either argumentatively or morally judgmentally more explicitly is necessary. That is, as long as people do not expect social robots to be capable of moral judgment and to intervene in these scenarios. This, however, might certainly change in the future upon further exposure to and development of social robots.

5.1.2 Positive and Negative Affect

Contrary to what I expected (H1a), there were no significant differences in change of affect between the three conditions. So, the conditions where the robot Pepper intervened more strongly (argumentative and morally judgmental) did not lead to a more positive increase in positive affect. Instead, for the positive affect subscale, there were no significant changes at all, both for the target person as well as the bystander. So, I have to reject H1a. What I did find, however, is support for H1b. There was a significant increase in negative affect for both bystander and target. The combination of these findings might be explained by the "negativity bias" that describes the phenomenon that negative events impact us more than positive or neutral events do (Ito et al., 1998), for example, in regards to emotion (Cacioppo and

Gardner, 1999). Dejonckheere et al. (2021) found that personally relevant negative events lead to a stronger bipolarity between positive and negative affect so that negative emotions are felt more strongly and positive affect is reduced. As the qualitative interview suggests, where most people directly reported the negative incident, participants of this study did, in fact, perceive this incident as quite negative and upsetting, underpinning the finding by Dejonckheere et al. (2021).

Interestingly, there were no significant differences between the two positions, target and bystander. I had assumed (H1c) that the targets would feel more negative than the bystanders due to the personal relevance and personal attack directed at them. This would also align with other research suggesting different perceptions based on how personally relevant the attack is (Brauer and Chekroun, 2005). However, it seems that even though bystanders were not the offence's target, the negative event would still elicit stronger negative emotions than before the intervention. The negative PANAS subscale asks for attributes such as being "upset", feeling "guilty", or "ashamed" (Watson et al., 1988). It could be that a different subset of negative affect was particularly elicited for bystanders, e.g. feeling guilty or ashamed for not having said anything. This is in line with comments by the bystanders in the interview, where they often mentioned how overwhelmed they were and how bad they felt for not having said anything. This matches with research by Hortensius and De Gelder (2018), who found critical situations to be emotionally stressful not only for the victims but also for the bystanders.

Visual inspection (see Figure 4.2) suggests, however, that the second condition (the argumentative one) had the highest impact on the increase in negative affect for the bystander position. This could be connected to most people in the second condition (compared to the other conditions) noticing that Pepper had intervened, potentially being more aware of the incident than the other conditions. In this condition, most bystanders had laughed or simply looked at the rest of the team, mirroring their insecurity. However, I see this trend only for the bystander position. For the target position, there is an increase in negative affect for all conditions - despite them reporting similar levels of awareness regarding Pepper's intervention. So, it seems that as a bystander, people would need to perceive a reaction by Pepper to be more aware of the assault and, through that, feel more negative about the encounter.

Similarly, for the positive affect subscale for the target position, there is a trend of an increase in the second condition. Despite this not being significant, it might suggest that noticing Pepper's intervention could lead to some changes in perception in the target person. Perhaps, upon experiencing Pepper's intervention, target people also experienced increased positive emotions, such as being excited or enthusiastic about Pepper's response or feeling more proud about themselves as a coping mechanism. See the following section (5.1.3) for a more detailed discussion.

Overall, however, no significant differences between the groups were found, which might be due to the relatively large variance inherent to the study design. Increasing the group sizes might already make the results more precise.

5.1.3 Self-Esteem

I have to reject H2a in that people will have a more positive change rate in self-esteem when the robot repairs the violation (the argumentative and morally judgmental conditions). The hypothesis meant, that there might well be a decrease in self-esteem. However, I expected self-esteem to more positively change from the pre-assessment to the post-assessment in the argumentative and morally judgmental condition compared to the avoidant condition. However, this change might still be negative.

Our results found a significant increase in all conditions for both the target and bystander positions. This is quite interesting as other research has found sexist encounters to reduce self-esteem in the affected people, both for hostile encounters (Swim and Hyers, 1999; Wippermann, 2022) as well as friendly sexist teasing (Hack et al., 2020). Potentially being in this situation and having successfully endured it led the people being the target of the sexist comment to more highly regard their self-esteem. This could be in line with other research that showed that people having intervened in this kind of situation did have a boost in self-esteem (Kaiser and Miller, 2004; Sabbagh et al., 2010).

While most of the participants in this study did not directly intervene and instead laughed (likely out of shock), many of them made eye contact with the rest of the team, accusingly stared at the confederate, initiated a conversation with Pepper, or directly continued the game. Potentially, having resorted to these reactions made them evaluate their self-esteem as higher afterwards.

Looking at the other team members and seeing their surprise and anger resonate might have made them feel understood and supported, as Gupta and Rathore (2021) have found in their research on support groups. This can be supported by statements participants made in the qualitative interview, such as "It was really helpful that VP3 [the bystander] and me had exchanged glances. This way, I felt that I am not the only person experiencing it this way". Having moved on by interacting with Pepper or continuing the game themselves might have given the participants the feeling that they had managed the situation well and did not lose sight of the task, empowering their self-esteem. Staring at the confederate might have given them the feeling that even though they did not directly address the offence, they still had let the confederate feel their anger. As Emler (2001) writes, people's perception has a crucial influence on their self-esteem, so, however they assess the sexist encounter might impact their self-esteem. So, if they felt like they had intervened or reacted in any other way to the incident, perhaps this positively influenced their self-esteem.

Another common reaction to conflict is disengagement (Laursen et al., 2001), where people distance themselves from the conflict. Swim and Stangor (1998) found that a higher disengagement correlates with higher self-esteem. So, potentially, people's reactions, such as directly continuing the game or not reacting to the incident at all,

are rather signs of disengagement and, therefore, explain the increase in self-esteem. This would also align with some target people's statements that they were quite used to these kinds of remarks and did not bother that much anymore.
Another possibility is that the group was relatively resilient, as resilience has been found to buffer the impact of sexism on self-esteem (Murphy Brien, 2023).

Regarding RQ3, how bystander and target differ, the slight significant increase (similar to that of the target position) that was seen in bystanders, on first look, seems surprising as there was no direct attack towards them personally. However, they might have felt empathy towards the victim (García-Ramírez, 2016), making it personal to them. And then, potentially for the same reasons as highlighted above, the bystanders might have thought of themselves as successfully having managed the situation, increasing self-esteem. Or, perhaps, as most of the bystanders did not *actively* react to the incident but simply laughed or looked at the others, this might also be a sign of disengagement. This then might also be connected with higher self-esteem (Swim and Stangor, 1998), potentially as a means of coping with the situation.

5.1.4 Social Attribution of Pepper

I mostly have to reject H3a, which supposes that the morally judgmental response would be perceived as more social than the argumentative response and that the argumentative response is more social than the avoidant one. For almost all subscales of the Robotic Social Attribution Scale, there was a significant pre-post difference: for both the targets and the bystanders, there was a significant increase in warmth and competence ratings, and for the targets, there was a significant decrease in the discomfort ratings. However, there were no significant differences between the conditions.

Following Paetzel et al. (2020), being confronted with a robot leads to a decrease in eeriness and an increase in competence. This is in line with our findings and seems stronger than the differences elicited through different reactions to the sexist comment by the robot. Of course, the robot's programming was the same for all conditions apart from the reaction to the sexist comment. As this reaction was only a fraction of the overall interaction with the robot, this difference may not be strong enough. Potentially, generally interacting with a social robot had a more substantial impact on the social ratings of the robot. Another point is that most people from the avoidant condition did not realise that the robot had said something about the incident, and quite a few target people from the morally judgmental condition also did not. Therefore, the lack of difference between the conditions might also be attributable to this difference in awareness of Pepper's intervention.

Although barely ($p = 0.053$) not significant, I can visually detect an interesting interaction effect in the warmth scale for the bystander, in that the values of the morally judgmental condition seem to increase more strongly than of the avoidant condition (see Figure 4.4). Despite this not being significant, I want to explore it a bit more as there are parallels to some qualitative findings. For one, the most common response for bystanders was to laugh, underpinning the warmth perception. Somewhat conflicting, most bystanders in the morally judgmental condition expressed feelings such as anger or irritation. Perhaps feeling anger and empathy towards the target in this situation might have led them also to experience more warmth regarding Pepper upon its response. While laughing and being angry somewhat appear to be mutually exclusive, the laugh could also have been a shocked laugh. Or perhaps these assessments came from different people, adding up to the increase in warmth. Ultimately, this interaction might have been a primary driver for the significant pre-post effect witnessed.

Reeder et al. (2002) found that morality is highly connected with warmth. Considering that the morally judgmental condition clearly judged the sexist comment to be inappropriate, this finding seems to make sense. Interestingly, however, the results did not show the same effect for the target position. Potentially, being the direct target of the sexist comment overrules whatever direct response was made by the robot, again coming from experiencing a norm violation differently when being directly affected (Brauer and Chekroun, 2005). I can also see this in the response by target people, where, compared to the bystander, less than half laughed at Pepper's response. The bystander person, in turn, as a person not directly affected (García-Ramírez, 2016), might view things differently.

Another explanation might be that in the morally judgmental condition, almost half of the bystanders thought that the sexist comment might not have been authentic, which is a lot more than in both other conditions. Viewing it under this light, potentially they saw the interaction mostly as funny, hence explaining the strong increase in warmth. This is congruent with comments that they thought the confederate was merely probing the robot on how it would react to such a critical comment.

Generally, there is a trend for the bystander position that the morally judgmental condition is seen as most competent, possessing the most warmth but eliciting the most discomfort. This would also be in line with findings by Winkle et al. (2021) where the aggressive condition, which might somewhat be comparable to our morally judgmental condition, was the only one where the interest in robots by boys did not decrease. As most people in the bystander position identified as male, this might also be attributable to the bystander position. This is also in line with Jackson et al. (2020), who found that men preferred aggressive robots of their own gender. Most of the predominantly male bystanders described Pepper as male. Therefore, this matches and might explain why I find (insignificantly) higher ratings for all subscales for the mostly male bystander position compared to assessments by the target position.

In contrast to Winkle et al. (2021), I do, however, see apparent increases in the social ratings of the robot. This might be because the target group is not school kids but already university students. Winkle et al. (2021) had identified that the boredom the school kids experienced might have led to this reduction in ratings.

Students participating in my study might not have been as bored, considering they participated voluntarily.

So, while there were no significant effects, interesting trends can be seen that point towards different robot perceptions of the bystander position. Expanding the sample size would potentially provide more insights here.

The only distinction going in this direction for the target person is that the competence ratings of the robot went up the most (again, not significant) for the argumentative condition. This would make sense as being able to argue properly should feed into the perception of competence, whereas morally judging a situation might not necessarily be seen as more competent than when rationally countering an argument.

Interestingly, discomfort was generally rated the lowest, warmth was in the middle, and competence took the top of the scale for both the bystander and the target. This suggests that there was generally only little discomfort felt towards Pepper, portraying an openness to interact with social robots – at least with the demographic tested in this study. However, I can also see that most participants did not directly address the robot but instead waited for the robot to address them. This could have been because I did not tell the participants how exactly they should address Pepper, as I was curious about how they would naturally navigate such a situation. Another finding regarding expectations towards the robot from the interviews was that a third of the people from the avoidant condition could not imagine Pepper reacting in an interpersonal way. This seems to be directly linked to not witnessing this behaviour from Pepper previously.

Of course, this interaction was only a short one and is not representative of when people are really confronted with working with social robots in the long run. However, these findings might provide additional insights as to how current students approach the topic of social robots. As a generation likely to interact with social robots more deeply, this is certainly insightful. Looking at the social attribution of robots is particularly relevant, as discomfort is connected with trust, which seems to be as a crucial topic for acceptance of social robots (Lewis et al., 2018).

It seems sensible that competence, out of the three scales, was rated the highest, as competence is what robots historically were built for: to be competent in whatever role they are to fill. A similar effect can be seen for algorithms in that people view them to be objective and competent (Helberger et al., 2020).

McKee et al. (2023) found that people perceive warmth and competence in AI (which is a crucial part of social robots) and that there seems to be a duality between warmth and competence: AI is perceived to be more competent when it operates independently from humans and as warmer when it operates in line with human interests. Potentially, the big increase in warmth stems from people witnessing their (human) interests being represented by the robot. What is interesting here is that both warmth and competence increased, so this duality does not seem to be mutually exclusive. Generally, looking at competence and warmth is interesting as those judgements predict how willing people are to interact with the system (McKee et al., 2023).

The present thesis shows that there is no need to worry that social robots will decrease in warmth and competence ratings when intervening in sexist encounters, which is a very promising result.

5.1.5 Team Conflict

I have to reject the hypothesis that there would be differences in team conflict perception based on the three conditions (H4a). Based on Jung et al. (2015), who had found a significant interaction effect when the robot intervened upon a personal violation, I had assumed that people would perceive the conflict to be higher when the robot repairs the violation. While most of our participants directly mentioned the conflict in the interview following the interaction, it seems that the different ways in which Pepper intervened did not necessarily elicit differences in conflict perception. This could point towards Pepper's interactions not being sufficiently distinct from another – at least in that regard –, or that it does not matter how the robot intervenes when it comes to conflict perception.

I can accept H4b that there is a greater relationship violation than task violation. This is as expected, as the experiment is focusing on a personal level, considering that the comment was sexist and not made on any valid basis.
I want to make you aware of an interesting trend in this regard, despite it not being significant. Namely, visual inspection (see Figure 4.5) shows that there might be a trend towards a difference between relationship and task conflict in how the assessment differs between the conditions. There was nearly no difference in the relationship conflict subscale between the different conditions. However, for the target person, there is a visible trend that task conflict perception was higher in the morally judgmental condition than in the argumentative condition. However, of course, there is much variance and no statistical significance. So, read the following interpretations with caution, as there is no statistical significance as the basis for it.

The visual inspection points to an interesting difference in contrast to Jung et al. (2015), who found differences mainly in the relationship conflict perception and not so much in the task conflict perception. My findings might have been different frome those by Jung et al., as their interpretation of "task violation" was that the confederate told the other participant, "Let's not use this one. Use this.". This does not necessarily sound like a *violation*, considering that it is acceptable and helpful to suggest another move in a game. Participants' confused laughter and looks upon the robot's intervention speak for this. They perhaps had not identified any task violation, and, therefore did not mention it in the questionnaires later on either.

Despite having a potential explanation for the differences in findings, the finding itself is quite interesting. One would not necessarily expect that Pepper intervening would lead to differences in task conflict perception. The scale asks questions regarding ideas within the team or how people approached their work. On the other hand, the personal conflict scale, as the name suggests, specifically asks about personal conflicts within the team. I thought this would be more applicable to the specific scenario of this study, considering that the sexist comment was a personal one. This is the case, as seen in H4b.

However, maybe Pepper's intervention did not matter regarding how big of a personal conflict there was, as people potentially did not feel that Pepper could change anything regarding those. However, considering that robots historically often were perceived more in terms of efficiency regarding the task rather than personal relations (Wang and Krumhuber, 2018), maybe people attribute more agency towards Pepper in regulating the task-related outcomes and conflicts rather than relationship conflicts. Therefore, based on Pepper's different responses, it might make sense that people perceived task conflict slightly (although insignificantly) differently.

One reason for these differences might be that people tend to punish norm violations (Molho et al., 2020). In this case they, potentially, viewed the argumentative intervention as having punished the offender enough. In contrast, in the morally judgmental condition, there may have been no closure, but instead, even further escalation so that people were more aware of the norm violation afterwards. It also makes sense that only the target position experienced it that strongly, as norm violation perceptions differ based on the level of personal affectedness (Brauer and Chekroun, 2005).

Interestingly, this goes in parallel with the finding of the qualitative interview that in the morally judgmental condition, more target people than in any other condition reported the incident only after explicitly being asked and not directly. Perhaps, not having had as much closure but instead rather being confronted with an escalated conflict, led people to hold back in the interview, perhaps to not engage further in the conflict. Interestingly, more people from the avoidant and morally judgmental condition than the argumentative one mentioned that intervening would not change anything or be helpful. This is interesting, as those were the conditions where Pepper actually intervened. So, perhaps, they did not experience the intervention as helpful after all. Or their prior experiences led to a certain disillusionment, which in turn may have played into not speaking about the incident.

But again, this is very hypothetical as it is only based on tendencies in the data and should not be taken too seriously. As it does align with other findings that gained significance, I thought it would be interesting to report and discuss this nonetheless. Looking at it with more data points could provide more insight.
So, to conclude, I can only say that Pepper's intervention seems to have little impact on people's perception of the conflict. However, the relationship conflict clearly was stronger than the task conflict.

5.1.6 Closeness to other Team Members

Regarding the closeness to the other team members, I have to reject H5a (that participants would feel closer to the robot in the argumentative and morally judgmental condition). I partly accept H5b (that the people perceive the robot differently than the human participants), as the target people viewed the robot significantly differently from the sexist offender but not from the bystander. I accept H5c (that the person making the sexist comment is perceived differently close than the other human participant). However, this again happened only for the target position and not the bystander.

So, generally, target participants rated the confederate as less close than their other team members. In contrast, Pepper was rated equally close or even slightly, though not significantly, closer (in the morally judgmental condition) than the human bystander. It makes sense that after an offence, the offended people would rate the offender as less close to them to reprimand the deviation from the group norm (DeMarco and Newheiser, 2018; Brauer and Chekroun, 2005). What is interesting to see is that in this particular situation, this leads to humans viewing the robot similarly to the bystander, which usually does not seem to be the case as humans tend to view social robots as not as socially close to them as other humans (Lanfranchi and Lemonnier, 2023).

Perhaps the norm violation through the sexist comment has led to a spontaneous in-group formation between the target, the robot and the bystander. The in-group bias describes the phenomenon that people view the in-group members more favourably than out-group members (Tajfel et al., 1971). Groups can be perceived based on seemingly arbitrary characteristics with the aim of devaluing out-group members to maintain higher self-esteem (Tajfel and Turner, 2004). Therefore, violating a norm might have triggered an in-group formation between the target person, the robot and the bystander against the offender – at least in the eyes of the target person. As a result, the sexist confederate was rated as less close, whereas Pepper was rated similarly to the other human group member.

So, it seems that having a "common enemy" leads people to assess a robot, otherwise considered less close, as similarly close to non-norm-violating group members. This is an interesting insight as this phenomenon might overrule other existing biases against social robots.
However, this effect could only be found this strong for the target person and not for the bystander. So, it seems that being personally affected decisively changes the perception of group members after a norm violation and extends to robotic agents. For people not personally affected by the norm violation, the findings by Lanfranchi and Lemonnier (2023) still seem to hold.

5.1.7 Team Member Perception

The same hypotheses I had for the feeling of closeness towards the individual team members I applied to the assessment of how good of a teammate the different team members were. The scale by Plum (2022) allowed me to assess team member perception from a different angle.

For the target position, in nearly all subscales I had significant differences in the assessment of the three team members in that the sexist confederate was rated significantly lower than the robot Pepper and the bystander (apart from the subscale "Knowing and Fulfilling Their Roles"). In contrast, from the bystander's point of view, I did not find any significant differences between the three team members in all four subscales that were evaluated. Here, I see similar results in the teammate subscales as in the team member closeness scale, suggesting these two constructs to correlate or overlap. Potentially, viewing people as good team members, makes them perceive themselves as closer to them. View section 5.1.6 for the detailed discussion of closeness.

Additionally, previous research by Plum (2022) found that in-group robots were considered less of a teammate than human in-group members. This is conceptually related to the finding by Lanfranchi and Lemonnier (2023) regarding perceived closeness with the robot, as mentioned in the previous section. Those findings are also congruent with previous research by Fraune (2020), who found that humans are valued more than robots, even if the robots outperform the humans.

Our research results now clearly contrast that. It seems that if people are the direct target of a sexist norm violation, this offence might overrule any usual perception regarding social robots. This is in line with research by Brauer and Chekroun (2005) in that people being more affected by a norm violation perceive this incident differently than others.

Interestingly, there are none of those significant effects for the bystander position, rendering their position less straightforward. Whereas for the target, the direct offence seems to dominate their perception, the picture for the bystander is more complex, as can be seen when visually exploring the results of the team member perception subscales (Figures 4.7 and 4.8). The qualitative interview suggests that bystanders experienced various emotions, from being overwhelmed and not knowing what to do, to blaming themselves for not having said or done anything during the intervention. Here, more data would be needed to understand the bystander's position better.

I would like to highlight two more findings for these subscales. For one, there was a significant difference between the argumentative and morally judgmental condition in the "Subjugating Individual Needs for Group Needs" subscale. Team members were assessed significantly higher in the argumentative condition than in the morally judgmental condition. Although not significant, I found a similar trend for the argumentative condition, which was also ranked higher than the avoidant one. A similar trend can also be detected for the "Knowing and Fulfilling their Roles" graph. Here, however, the sexist confederate was rated slightly, but insignificantly, higher than the other two team members.

One reason might be that most target people had actually noticed Pepper's intervention in the argumentative condition. Potentially, only noticing the intervention is already strong enough to shift up the group member assessments. Here, the explanation could be similar to the one I raised in section 5.1.5 regarding team conflict perception

Perhaps, in the argumentative condition, the conflict was considered "solved", whereas the morally judgmental condition further escalated the problem. In the avoidant condition, most people had not noticed Pepper's intervention. Therefore, the conflict was lingering on here as well. This is in parallel with people not being as confident in the morally judgmental and avoidant condition, compared to the argumentative one, that countering sexism could have a beneficial effect. Perhaps how present the conflict was in people's minds caused them to become less confident and rather overwhelmed.

This is also somewhat mirrored in the findings of decreasing anger over the three conditions that was found for the target (83% for the avoidant condition, 58% for the argumentative condition and 38% for the morally judgmental condition). The percentage for the argumentative condition somewhat speaks against this interpretation. However, looking at emotional changes between the avoidant and argumentative conditions would support it. Potentially, in the morally judgmental condition, target people were, despite the perception of escalated conflict, still less angry, considering that there was a stronger acknowledgement of the sexist encounter.

More research is necessary to understand the role of anger in speaking up or feeling empowered. For the moment, it seems that the argumentative condition is the most promising concerning team member perception. This would be in line with findings by Winkle et al. (2021), who showed the argumentative condition to be the most well-received by people identifying as female.

Another interesting finding was that targets ranked the sexist confederate significantly worse in the avoidant condition than in the argumentative and morally judgmental condition in the subscale "Social Interaction" (e.g. "My team member and I could work well together", "I am happy, Person X and I were in the same team"). Here again, it seems that if Pepper does not sort out the conflict, people use the questionnaire to let out their frustration and penalize the norm violation by the sexist confederate.

For the bystander, I did not find that the sexist confederate was rated significantly lower than the other team members. This again might be linked to the discussion about "Social Attribution of Pepper" (see section 5.1.4). Here, I identified the gender of the bystander position as potentially being linked to rating the morally judgmental condition higher. Although the differences between the conditions are not significant in this subscale, the visual inspection (Figure 4.8) shows that the morally judgmental condition generally seems to be on top of all subscales. Perhaps there is an inclination to view everyone on the team more favourably if the robot intervenes more strongly. However, this effect is not significant and needs to be explored further with more data.

5.1.8 Pepper's Gender

As the gender people perceive a robot to have can highly impact people's perception of the robot (Eyssel and Kuchenbrandt, 2012), it was essential for me not to label Pepper as having a gender. The reason was that gendering technology might reinforce stereotypes (West et al., 2019; Galatolo et al., 2022). However, people sometimes still subconsciously assign a gender to gender-ambiguous agents (Sutton, 2020). Therefore, I considered it important to assess which potential effect this might have on the study results. I found that most people subconsciously used the male pronoun to describe Pepper. However, when asked which gender they would assign the robot, most people said Pepper was neutral. Potentially, this is because of the social desirability bias (Grimm, 2010), as people know that robots are not human and might, therefore, believe they should say that the robot has no gender.

Implicitly, though, there could still be a bias to view Pepper as male. However, in the German language, the article to the word robot is male ("*der* Roboter") (Duden, 2024). Therefore, it is likely that many people used the male pronoun simply because they transferred the gender from the word robot to Pepper. This is also how 12% of the participants described it. A study by Shin and Kim (2007), done in Korea, showed a similar trend. The least people in my study thought Pepper was female, which would be in line with stereotypes of not connecting women with technology (West et al., 2019). Interestingly, opinions diverge when it comes to whether the voice or name now is male or female, as these were used as reasons arguing for both genders.

The relationship between Pepper's gender and the different conditions seems quite complex. Regarding the *implicit* labelling of Pepper through the use of pronouns, it seems that in the morally judgmental condition compared to the other two conditions, it is *not* the male pronoun that is predominant. Instead, using the female, male, or no pronoun balances each other out. Of course, I cannot say anything about causality here. Perhaps people in this condition just deviated from the people of the other conditions.

There could, however, also be a link: Women stereotypically are believed to be more emotionally expressive than men (Shields, 2002). Perhaps witnessing Pepper using a morally judgmental response let more people take on the perception of Pepper being female. This somewhat aligns with people's *explicit* labelling of Pepper's gender as well. In the avoidant condition, most people labelled Pepper as male, followed by neutral, whereas nearly no one viewed Pepper as female. In the argumentative condition, most people viewed Pepper as neutral, closely followed by male. In the morally judgmental condition, this turned, and more people than in the avoidant condition saw Pepper as neutral, closely followed by female. Perhaps the extent to which Pepper intervened changed people's default male perspective to viewing Pepper's gender differently. One person even mentioned this explicitly in the interview.

Hearing other participants talk about Pepper with one pronoun or the other might have influenced their perception as well. However, when discussing this in the interview, they often mentioned how confused they were when hearing the other person talk about Pepper with "he" or "she", respectively.

Viewing Pepper's gender a certain way might have an impact on other related measures (Galatolo et al., 2022). However, more research is necessary to understand this complex relationship.

5.1.9 General

As both qualitative and quantitative results show, this study made an impression on the participants. Independent of condition or position (target/bystander), almost all immediately reported the incident. Also, the significant pre-post evaluations speak for the impact of the experiment. There was a big need to discuss the study in the debriefing and to hear how other people had reacted. The feeling of relief that it was just an act was omnipresent. Many, particularly target people, were very agitated; some even started crying during the interview. While I had anticipated, prior to conducting the study, that this might be a challenging experience for the participants, I admittedly underestimated the extent to which participants would react emotionally to the incident. This should be taken into account when trying to replicate the study.

Most people, normally identifying as relatively progressive and feminist, as seen by the participants' political stance, had to realise that they had not taken up the courage to intervene.

At this point, I should add that however one reacts to a sexist encounter is valid (Kaiser and Miller, 2004). There are a multitude of factors that play into how people perceive and react to sexist comments. Even though countering sexism might be a good way to get offenders to understand that their sexist comments are inappropriate and potentially change offenders' future behaviour (Neoh et al., 2023), one should not blame women for other responses. Instead, research suggests that to move to a more gender-equal world, perspectives and reactions of women should be validated (National Academies of Sciences et al., 2018). Additionally, the focus should move towards how perpetrators should change their behaviour rather than criticising women's responses towards sexist encounters.

Despite this being the case, many participants were in conflict with their reaction. It seems they experienced a cognitive dissonance (Harmon-Jones and Mills, 2019) regarding how they had reacted and how they thought they would react in these situations. Especially male bystanders struggled in the interview as they had never experienced such an openly sexist situation with the victim being present. This was also reflected in the big difference in reactions, where more than double of bystander people looked at the other team members compared to target people, potentially in search of help or out of confusion about what to do. One male bystander wrote down Pepper's interventions to better prepare for future sexist encounters. Another bystander openly reflected that he now knew how *not* to react and how he could improve in the future. Some even thanked me for the experience in such a safe environment. So, it seems that no matter Pepper's actual intervention, participants already had a major takeaway regarding their own response.

Interesting to highlight here, still, is the minor number of people choosing to directly confront the confederate, especially considering that it was "just" a lab study. For none of them the outcome of this study had any impact on their future in the sense that they really had to perform well. So, it was not like they had much to loose which usually makes it harder to speak up (J. Nicole and Stewart, 2004).

However, one should consider the effect of social comparison (Hogg, 2000) and competitive pressure when people are told that their group performance is evaluated. Perhaps they did not want to perform poorly compared to other groups by further disrupting the group work. At least this is reflected in the answers they gave, like not wanting to "make a fuss about it". Other responses reflected their limited time as a reason for not intervening. However, considering that they had ten minutes and all but two teams managed to finish in time, this, at least rationally, does not seem to be a valid reason. However, time pressure might, of course, change people's general approach to the task, and their behaviour might have been different in another setting.

Other reasons people put forth were that they were not sure whether the sexist confederate really had meant it as an offence or whether he was only joking. Potentially, downplaying the intention of the offender allowed the participants to be more at ease. Bellezza et al. (2014), for example, found that if a norm violation is unintentional, it does not pose as much of a threat to the group. So, this seems to have been another way to downplay the criticality of the situation and reduce people's cognitive dissonance for not having said something.
Other possibilities are that it simply is uncomfortable to confront someone. Especially women are trained not to be disagreeable (Hamid et al., 2010). This subconsciously might lead to many women not speaking up.

Putting all this aside, it remains that even in a setting with objectively not much to lose, potentially much less to lose than in any real-life scenario, considering that the study was video-monitored, and despite the majority of participants identifying as progressive, only very few people decided to intervene. Moreover, even if there are reasons such as time pressure, this study's findings could still be generalisable to real-life encounters outside of the lab, as it is always possible to come up with reasons as to why one could not intervene.

So, it seems that there still is a lot more work to be done when wanting to counter sexism through active intervention. As this study shows, a social robot intervening might, in fact, really support both offended people as well as bystanders. However, one could also question whether active confrontation necessarily is the way to go.

Furthermore, of course, intervention through technology should not develop to the other extreme in that people start relying on that technology to solve their interpersonal conflicts and, in turn, no longer confront any conflicts on their own. However, considering the current state of technology, this worry seems rather far down the line.

5.2 Limitations

One main limitation of this study is the variance I have witnessed. For one, this was achieved through the study design. The goal was to create a realistic scenario where people did not suspect something to be off about the sexist comment. Therefore, the sexist comment was always spoken after the target person said something. If the target person did not propose anything herself, the confederate or Pepper asked her for her suggestion. Being asked or proposing something yourself are very different things and might, therefore, change how people perceive a sexist comment, leading to an increase in variance.

Additionally, as I wanted to see how the study participants reacted to the sexist comment, providing them with some freedom after it was spoken was necesssary. This led to some variance from people not saying anything so that the confederate had to guide the group through, over people laughing at Pepper's interaction and simply continuing with the game, to other participants taking over the lead and almost ignoring the confederate. All of this may have impacted people's assessment of the situation and their answers to the scale, explaining the huge variances I have seen.

Another limitation is the noticeable difference in how the three men portraying the confederate acted in their role. It was a challenge for all three of them to be so openly sexist and bear the following awkward situation. Therefore, I decided that it was more important that the sexist comment was spoken believably so that participants would not question the authenticity of the comment. However, every one of them interpreted their role slightly differently, so there might also be differences in the perception of the experiment based on which confederate the participants had experienced. However, I used stratified randomisation, so at least any variances that have arisen as a result are equally distributed across the three conditions.

Another limitation potentially influencing the results is that Pepper, unfortunately, sometimes lost connection to the internet, making it impossible to steer the robot. I then had to reboot Pepper. A standard procedure for this was to reboot the robot right before starting the game so that the interaction mostly went fluently. However, sometimes Pepper crashed more than once. Restarting Pepper might influence how people perceive Pepper's sociability. I included this as a covariate in the relevant scales. However, there might still be influences beyond my perception.

In case one participant spontaneously did not show up, I had another confederate as a backup who would jump in, as the target or bystander. This confederate was trained regarding how to react in the respective situations and was told not to influence the group dynamics to get a pure response of the other participant. However, this of course, might have also influenced the whole experimental perception.

In order to keep the experience of the participants smooth, the interviews at the end were done in parallel in two different rooms. For room capacity reasons, I did one interview in the main lab and the other in the adjacent room. Of course, both participants now had different surroundings, which might subconsciously impact people's answers. Additionally, the person remaining in the main lab was still in the same robot as Pepper, who, at this point, was idle but still running. When answering questions about the robot, being in one room with Pepper might have made participants answer them differently than not being in the same room, considering the findings by Nass et al. (1999) that people rate a computer on a different computer in another room differently than when on the computer itself. Merely the presence of Pepper might have elicited some subconscious social desirability bias when talking about the robot (Grimm, 2010).

Another limitation potentially leading to some variance is that some people were non-native speakers of the German language and might have had more difficulties understanding the complex team relations in this fast-paced scenario. This is also in line with some participants stating that they did not hear the confederate's comment properly. However, I decided to keep these people in the assessment as it reflects how real-life situations might happen.

The groups were generally quite small, ranging from ten to thirteen participants. Therefore, group size certainly is a limitation. For more robust results, more participants are needed.

Apart from these limitations leading to variance in the results, there are also some limitations in the general design of the study. For one, the sexist comment had to be somewhat obvious to work as a manipulation. However, sexism today often is prevalent in more unobtrusive ways, for example women being ignored or interrupted when they are trying to add to the conversation (Wippermann, 2022). However, having a scenario with such a direct sexist comment, as in my study, should still be able to provide realistic insights regarding people's perceptions of interventions of robots regarding direct sexist encounters.

Another limitation is the time pressure, as mentioned in section 5.1.9. Time pressure was added as a means of reproducing the study by Jung et al. (2015) to add as an additional stress component and to increase the impact of the violation. It was also introduced to keep participants from wandering too far off the track of what they would ask Pepper – considering that it was a Wizard-of-Oz study. While almost all teams finished in time, being in the mindset of being under time pressure might move people into a different state of mind. In order to have situations that might generalise better to everyday sexist encounters, it might be worth exploring the same experiment without the time pressure.

Regarding the evaluation of the study: For one, the qualitative analysis was done by only one person, considering that it was a master's thesis. Realistically, one would want to have at least one more person involved in the analysis to compute the inter-rater reliability (Hallgren, 2012). This would be done to ensure objectivity or at least consistency in the assessment between different people evaluating the observation.

Secondly, in the team conflict scale, some participants were not sure whether to answer in numbers as to how many conflicts exactly they have had in the team or whether it was in line with the description of the numbers ("very little" or "very much"). This may have skewed the analysis of this scale a bit.

And lastly, some general remarks. The study was a lab study and, therefore, might lack generalisability to real-life scenarios (Brunswik, 1955). Additionally, most of the study participants were students from a university, which again might lead to limited generalisability (Sears, 1986). Also, participation was voluntary and compensated with 15€, potentially attracting a specific subset of people, again limiting the generalisability of the study's findings.

Despite all these limitations, the effects witnessed still inform how people react to a robot intervening in sexist encounters.

5.3 Future Work

Considering the limitation of having much variance, it would be interesting to continue the study with more study participants to gain greater power and be able to tell more conclusive results. To achieve this, it might also make sense to reduce the complexity of the study design. For example, it might be an option to keep the bystander position constant and have a confederate always act in this position. Alternatively, to at least minimally reduce variance, one could have only *male* people in the bystander position, considering that gender might influence people's perception of the situation, by either being affected by sexist comments or not.

Considering the limitation regarding sexist comments typically being less direct, it would be interesting to assess how people's perception of the intervention would change in case of less direct sexism, for example, a male confederate not letting a woman speak but constantly interrupting her. Of course, it might be more challenging to make people believe that a robot can realistically detect this kind of sexism. However, this scenario could provide interesting insights regarding more complex cases of sexism.

Other dimensions to extend this to are any other kinds of discriminatory situations, such as racist insults or comments directed against people identifying as queer. It would be interesting to see whether there are parallels in perception between different marginalised communities or whether different sub-patterns of reactions and empowerment emerge.

Another point mentioned in the limitations is the time pressure people experienced during the study. Potentially, it would be helpful to remove the time-pressure aspect entirely to be able to generalise results better to real-life encounters. For this, Pepper would need more comprehensive programming to intercept potentially critical questions directed at the robot. The confederate should further suffice to influence the team's progress in the game unobtrusively.

In order to reduce complexity, I, in this study, only looked at cis-people, i.e. people identifying with the sex they were born with Enke (2012). However, in line with current best practices (Winkle et al., 2023), it would be particularly insightful how trans people or anyone not identifying as one of the two binary "male" and "female" genders, act in these concrete sexist encounters.

Chapter 6

Conclusion

The present study aimed to explore how a social robot could effectively intervene in sexist encounters, providing support to both the victim and bystander. Given the pervasive nature of sexism in society and the potential for technology to perpetuate biases, understanding strategies to address sexism is crucial for promoting gender equity. For this purpose, I conducted a mixed-methods laboratory study involving a group scenario in which participants played the game *Mastermind* alongside the social robot Pepper. In response to a sexist comment made by a male confederate, Pepper intervened in one of three ways: avoidant, argumentative, or morally judgmental.

The findings revealed that exposure to a sexist comment significantly impacted participants, eliciting heightened negative emotions. People being the victim of the comment rated the sexist confederate significantly worse than the bystander and Pepper. Contrary to previous studies, which often found robots to be ranked lower than humans, our results suggest that Pepper's intervention in conflict situations may enhance perceptions of its teamwork capabilities, potentially comparable even to human counterparts. This finding may have significant implications for the acceptance of robots, suggesting a push for further exploration into developing morally competent robotic systems.

Avoidant interventions often were ineffective as they did not name sexism specifically, leading to participants frequently failing to recognise Pepper's intervention. Those participants were more likely to resume the game than to actively challenge the confederate by looking at him or interacting with Pepper as a response to Pepper's intervention. This suggests that an avoidant intervention might not be enough to support the victim and truly counter sexism. Instead, more direct forms of intervention, such as argumentative or morally judgmental responses, should be preferred. There are tendencies that an argumentative response leads to better overall perceptions of teamwork, whereas a morally judgmental response risks escalating conflicts. More research with an increased sample size is necessary to assess these trends more clearly.

In conclusion, this study demonstrates the potential for social robots to provide meaningful support in sexist encounters, benefiting both victims and bystanders. Moreover, the robot's intervention spurred introspection, resulting in a learning opportunity for individuals to reflect on their own reactions. Future research should look at intervention options through social robots more closely to harness their potential.

Bibliography

Ajoudani, A., Zanchettin, A. M., Ivaldi, S., Albu-Schäffer, A., Kosuge, K., and Khatib, O. (2018). Progress and prospects of the human–robot collaboration. *Autonomous Robots*, 42:957–975.

Aldebaran (2023a). Choreographe Suite.

Aldebaran (2023b). Pepper the humanoid and programmable robot | Aldebaran.

Anderson, E. S. (1999). What is the point of equality? *Ethics*, 109(2):287–337.

Axelrod, L. and Hone, K. (2005). E-motional advantage: performance and satisfaction gains with affective computing. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pages 1192–1195.

Ayres, M. M., Friedman, C. K., and Leaper, C. (2009). Individual and situational factors related to young women's likelihood of confronting sexism in their everyday lives. *Sex Roles*, 61(7-8):449–460.

Bartneck, C., Van Der Hoek, M., Mubin, O., and Al Mahmud, A. (2007). " daisy, daisy, give me your answer do!" switching off a robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 217–222.

Bellezza, S., Gino, F., and Keinan, A. (2014). The red sneakers effect: Inferring status and competence from signals of nonconformity. *Journal of consumer research*, 41(1):35–54.

Berkovits, I., Hancock, G. R., and Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: Relative robustness to sphericity and normality violations. *Educational and Psychological Measurement*, 60(6):877–892.

Blut, M., Wang, C., Wünderlich, N. V., and Brock, C. (2021). Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other ai. *Journal of the Academy of Marketing Science*, 49:632–658.

Box, G. E. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4):317–346.

Brauer, M. and Chekroun, P. (2005). The relationship between perceived violation of social norms and social control: Situational factors influencing the reaction to deviance. *Journal of Applied Social Psychology*, 35(7):1519–1539.

Breazeal, C., Dautenhahn, K., and Kanda, T. (2016). *Social Robotics*, pages 1935–1972. Springer International Publishing, Cham.

Briggs, G. M. and Scheutz, M. (2015). “sorry, i can't do that”: Developing mechanisms to appropriately reject directives in human-robot interactions. In *2015 AAAI fall symposium series*.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological review*, 62(3):193.

Burgoon, J. K. (2015). Expectancy violations theory. *The international encyclopedia of interpersonal communication*, pages 1–9.

Cacioppo, J. T. and Gardner, W. L. (1999). Emotion. *Annual review of psychology*, 50(1):191–214.

Carpenter, J., Davis, J. M., Erwin-Stewart, N., Lee, T. R., Bransford, J. D., and Vye, N. (2009). Gender representation and humanoid robots designed for domestic use. *International Journal of Social Robotics*, 1:261–265.

Carpinella, C. M., Wyman, A. B., Perez, M. A., and 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*, pages 254–262.

Churchland, P. S. (2011). *Braintrust: What neuroscience tells us about morality*. Princeton University Press.

Ciarrochi, J. and Bilich, L. (2006). Acceptance and commitment therapy. measures package. *Unpublished manuscript, University of Wollongong, Wollongong, Australia*.

Corp, I. (2023). Ibm spss statistics for macintosh, version 29.0.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Dardenne, B., Dumont, M., and Bollier, T. (2007). Insidious dangers of benevolent sexism: consequences for women's performance. *Journal of personality and social psychology*, 93(5):764.

Dejonckheere, E., Mestdagh, M., Verdonck, S., Lafit, G., Ceulemans, E., Bastian, B., and Kalokerinos, E. K. (2021). The relation between positive and negative affect becomes more negative in response to personally relevant events. *Emotion*, 21(2):326.

DeMarco, T. C. and Newheiser, A.-K. (2018). Coping with group members who insult the in-group. *Social Psychological and Personality Science*, 9(2):234–244.

Duden (2024). Roboter, der.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and autonomous systems*, 42(3-4):177–190.

Edwards, A., Edwards, C., Westerman, D., and Spence, P. R. (2019). Initial expectations, interactions, and beyond with social robots. *Computers in Human Behavior*, 90:308–314.

Edwards, C. and Myers, S. A. (2007). Perceived instructor credibility as a function of instructor aggressive communication. *Communication Research Reports*, 24(1):47–53.

Emler, N. (2001). Self-esteem. *The costs and causes of low self-worth*.

Enke, A. F. (2012). The education of little cis. *Transfeminist perspectives in and beyond transgender or gender studies*, pages 60–77.

Esterwood, C. and Robert, L. P. (2020). Human robot team design. In *Proceedings of the 8th international conference on human-agent interaction*, pages 251–253.

Eyssel, F. and Hegel, F. (2012). (s) he's got the look: Gender stereotyping of robots 1. *Journal of Applied Social Psychology*, 42(9):2213–2230.

Eyssel, F. and Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, 51(4):724–731.

Ferreira, M. and Fletcher, S. (2021). *The 21st Century Industrial Robot: When Tools Become Collaborators*. Intelligent Systems, Control and Automation: Science and Engineering. Springer International Publishing.

Ferring, D. and Filipp, S.-H. (1996). Messung des selbstwertgefühls: befunde zu reliabilität, validität und stabilität der rosenberg-skala. *Diagnostica-Gottingen-*, 42:284–292.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. sage.

Fraune, M. R. (2020). Our robots, our team: Robot anthropomorphism moderates group effects in human–robot teams. *Frontiers in psychology*, 11:540167.

FurhatRobotics (2023). Furhat robotics.

Galatolo, A., Melsión, G. I., Leite, I., and 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*, pages 1–15.

Garcha, D., Geiskkovitch, D., Thiessen, R., Prentice, S., Fischer, K., and Young, J. (2023). Face to face with a sexist robot: Investigating how women react to sexist robot behaviors. *International Journal of Social Robotics*, pages 1–20.

García-Ramírez, G. M. (2016). Victim, perpetrator and bystander perspectives: Vatiations in language usage, empathy and violence sensitivity.

Geisser, S. and Greenhouse, S. W. (1958). An extension of box's results on the use of the f distribution in multivariate analysis. *The Annals of Mathematical Statistics*, 29(3):885–891.

Gheaus, A. and Robeyns, I. (2011). Equality-promoting parental leave. *Journal of Social Philosophy*, 42(2):173–191.

Glick, P. and Fiske, S. T. (1997). Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of women quarterly*, 21(1):119–135.

Gosten, S. (2023). mastermind. https://github.com/tabalugabu/mastermind/tree/alt_solution.

Grimm, P. (2010). Social desirability bias. *Wiley international encyclopedia of marketing*.

Groom, V. and Nass, C. (2007). Can robots be teammates?: Benchmarks in human–robot teams. *Interaction studies*, 8(3):483–500.

Gupta, S. and Rathore, H. S. (2021). Socio-economic and political empowerment through self help groups intervention: A study from bilaspur, chhattisgarh, india. *Journal of Public Affairs*, 21(1):e2143.

Hack, T., Garcia, A. L., Goodfriend, W., Habashi, M. M., and Hoover, A. E. (2020). When it is not so funny: prevalence of friendly sexist teasing and consequences to gender self-esteem. *Psychological Reports*, 123(5):1934–1965.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

Hamid, S., Johansson, E., and Rubenson, B. (2010). Security lies in obedience-voices of young women of a slum in pakistan. *BMC public health*, 10:1–7.

Harmon-Jones, E. and Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory.

Helberger, N., Araujo, T., and de Vreese, C. H. (2020). Who is the fairest of them all? public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, 39:105456.

Hogg, M. A. (2000). Social identity and social comparison. In *Handbook of social comparison: Theory and research*, pages 401–421. Springer.

Hortensius, R. and De Gelder, B. (2018). From empathy to apathy: The bystander effect revisited. *Current Directions in Psychological Science*, 27(4):249–256.

Howard, A. and Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24:1521–1536.

Inc., P. T. (2015). Collaborative data science.

Infante, D. A. (1987). *Arguing constructively*. Waveland Press.

Ito, T. A., Larsen, J. T., Smith, N. K., and Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of personality and social psychology*, 75(4):887.

J. Nicole, S. and Stewart, R. E. (2004). Confronting perpetrators of prejudice: The inhibitory effects of social costs. *Psychology of Women Quarterly*, 28(3):215–223.

Jackson, R. B., Williams, T., and 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*, pages 559–567.

Jung, M. F., Martelaro, N., and Hinds, P. J. (2015). Using robots to moderate team conflict: the case of repairing violations. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 229–236.

Kaiser, C. R. and Miller, C. T. (2004). A stress and coping perspective on confronting sexism. *Psychology of women quarterly*, 28(2):168–178.

Kieliszek, Z. et al. (2019). Future conflicts are inevitable: Causes of interpersonal conflicts according to immanuel kant and thomas r. malthus. *Philosophy and Cosmology*, 22(22):152–161.

Knuth, D. E. (1976). The computer as master mind. *Journal of Recreational Mathematics*, 9(1):1–6.

Korcz, Z. (2016). mastermind. `https://github.com/Calanthe/mastermind`.

Krohne, H. W., Egloff, B., Kohlmann, C.-W., Tausch, A., et al. (1996). Untersuchungen mit einer deutschen version der" positive and negative affect schedule"(panas). *Diagnostica-Gottingen-*, 42:139–156.

Kurfess, T. R. et al. (2005). *Robotics and automation handbook*, volume 414. CRC press Boca Raton, FL.

Kwon, M., Jung, M. F., and Knepper, R. A. (2016). Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 463–464. IEEE.

Lanfranchi, J.-B. and Lemonnier, S. (2023). The estimation of physical distances between oneself and a social robot: Am i as far from the robot as it is from me? *Europe's Journal of Psychology*, 19(3):299.

Latikka, R., Savela, N., Koivula, A., and Oksanen, A. (2021). Attitudes toward robots as equipment and coworkers and the impact of robot autonomy level. *International Journal of Social Robotics*, 13(7):1747–1759.

Laursen, B., Finkelstein, B. D., and Betts, N. T. (2001). A developmental meta-analysis of peer conflict resolution. *Developmental review*, 21(4):423–449.

Leiner, D. J. (2024). Sosci survey (version 3.5.00): [computer software].

Levene, H. (1960). Robust tests for equality of variances. *Contributions to probability and statistics*, pages 278–292.

Lewis, M., Sycara, K., and Walker, P. (2018). The role of trust in human-robot interaction. *Foundations of trusted autonomy*, pages 135–159.

Lopatovska, I., Brown, D., and Korshakova, E. (2022). Contextual perceptions of feminine-, masculine-and gender-ambiguous-sounding conversational agents. In *International Conference on Information*, pages 459–480. Springer.

Malle, B. F. and Scheutz, M. (2020). Moral competence in social robots. In *Machine ethics and robot ethics*, pages 225–230. Routledge.

Marcos-Pablos, S. and García-Peñalvo, F. J. (2022). Emotional intelligence in robotics: a scoping review. In *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence: The DITTET Collection 1*, pages 66–75. Springer.

Mast, M. S. (2005). The world according to men: It is hierarchical and stereotypical. *Sex Roles*, 53:919–924.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, 11(2):204–209.

Mayring, P. and Fenzl, T. (2019). *Qualitative inhaltsanalyse*. Springer.

McKee, K. R., Bai, X., and Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. *Iscience*, 26(8).

Molenberghs, P. (2013). The neuroscience of in-group bias. *Neuroscience & Biobehavioral Reviews*, 37(8):1530–1536.

Molho, C., Tybur, J. M., Van Lange, P. A., and Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature communications*, 11(1):3432.

Murphy, R. and Woods, D. D. (2009). Beyond asimov: The three laws of responsible robotics. *IEEE intelligent systems*, 24(4):14–20.

Murphy Brien, S. (2023). *Assessing the effects of sexism on self-esteem and resilience and resilience as a moderator*. PhD thesis, Dublin, National College of Ireland.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103.

Nass, C., Moon, Y., and Carney, P. (1999). Are people polite to computers? responses to computer-based interviewing systems 1. *Journal of applied social psychology*, 29(5):1093–1109.

National Academies of Sciences, E., Medicine, et al. (2018). Sexual harassment of women: climate, culture, and consequences in academic sciences, engineering, and medicine.

Neoh, M. J. Y., Teng, J. H., Setoh, P., and Esposito, G. (2023). Perceptions of sexism interact with perceived criticism on women's response to sexist remarks in different relationship types. *Scientific Reports*, 13(1):18393.

OpenAI (2023). OpenAI DevDay, Opening Keynote.

Paetzel, M., Perugia, G., and Castellano, G. (2020). The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pages 73–82.

pandas development team, T. (2024). pandas-dev/pandas: Pandas.

Plum, L. (2022). *Let's Team Up: Investigating the Perception of Robotic Teammates in Human-Robot-Dyads*. PhD thesis, RWTH Aachen University.

Profeta, P. (2020). *Gender equality and public policy: Measuring progress in Europe*. Cambridge University Press.

Pulliam-Moore, C. (2015). Google Photos identified black people as 'gorillas,' but racist software isn't new.

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., and Trafimow, D. (2002). Inferences about the morality of an aggressor: the role of perceived motive. *Journal of personality and social psychology*, 83(4):789.

Rheinheimer, D. C. and Penfield, D. A. (2001). The effects of type i error rate and power of the ancova f test and selected alternatives under nonnormality and variance heterogeneity. *The Journal of Experimental Education*, 69(4):373–391.

Rose, A. (2010). Are Face-Detection Cameras Racist?

Sabbagh, M., Hare, T., Wheelhouse, E., and McFarland, H. (2010). Self-silencing in response to sexist behavior: Exploring women's willingness to confront sexism. *The Pegasus Review: UCF Undergraduate Research Journal*, 4(2):2.

Savela, N., Kaakinen, M., Ellonen, N., and Oksanen, A. (2021). Sharing a work team with robots: The negative effect of robot co-workers on in-group identification with the work team. *Computers in human behavior*, 115:106585.

Schmader, T. and Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of personality and social psychology*, 85(3):440.

Seaborn, K. and Frank, A. (2022). What pronouns for pepper? a critical review of gender/ing in research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Seaborn, K., Miyake, N. P., Pennefather, P., and Otake-Matsuura, M. (2021). Voice in human–agent interaction: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–43.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of personality and social psychology*, 51(3):515.

Sen, A. (1993). Capability and well-being73. *The quality of life*, 30:270–293.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Shields, S. A. (2002). *Speaking from the heart: Gender and the social meaning of emotion*. Cambridge University Press.

Shin, N. and Kim, S. (2007). Learning about, from, and with robots: Students' perspectives. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 1040–1045. IEEE.

Song-Nichols, K. and Young, A. (2020). Gendered robots can change children's gender stereotyping. In *CogSci*.

Stone, W. L. (2018). *The history of robotics*. CRC Press Boca Raton, FL.

Sutton, S. J. (2020). Gender ambiguous, not genderless: Designing gender in voice user interfaces (vuis) with sensitivity. In *Proceedings of the 2nd conference on conversational user interfaces*, pages 1–8.

Swim, J. K. and Hyers, L. L. (1999). Excuse me—what did you just say?!: Women's public and private responses to sexist remarks. *Journal of experimental social psychology*, 35(1):68–88.

Swim, J. K., Hyers, L. L., Cohen, L. L., and Ferguson, M. J. (2001). Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of Social issues*, 57(1):31–53.

Swim, J. K. and Stangor, C. (1998). *Prejudice: The target's perspective*. Elsevier.

Tajfel, H., Billig, M. G., Bundy, R. P., and Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology*, 1(2):149–178.

Tajfel, H. and Turner, J. C. (2004). The social identity theory of intergroup behavior. In *Political psychology*, pages 276–293. Psychology Press.

Tavakol, M. and Dennick, R. (2011). Making sense of cronbach's alpha. *International journal of medical education*, 2:53.

Tay, B., Jung, Y., and 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.

Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, J. M. M., and Bernstein, A. (2021). Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

UN (2023a). Goal 5. Achieve gender equality and empower all women and girls.

UN (2023b). The Sustainable Develoment Agenda.

Vasey, M. W. and Thayer, J. F. (1987). The continuing problem of false positives in repeated measures anova in psychophysiology: A multivariate solution. *Psychophysiology*, 24(4):479–486.

Voiklis, J., Kim, B., Cusimano, C., and Malle, B. F. (2016). Moral judgments of human vs. robot agents. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 775–780. IEEE.

Wang, S. and Bunt, A. (2017). Surveying initiatives aimed at increasing female participation in computer science. Technical report.

Wang, X. and Krumhuber, E. G. (2018). Mind perception of robots varies with their economic versus social function. *Frontiers in psychology*, 9:344193.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063.

West, M., Kraut, R., and Ei Chew, H. (2019). I'd blush if i could: closing gender divides in digital skills through education.

Wilson, E. O. (2000). *Sociobiology: The new synthesis*. Harvard University Press.

Winkle, K., Jackson, R. B., Melsión, G. I., Brščić, D., Leite, I., and 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)*, pages 120–129. IEEE.

Winkle, K., Lagerstedt, E., Torre, I., and Offenwanger, A. (2023). 15 years of (who) man robot interaction: Reviewing the h in human-robot interaction. *ACM Transactions on Human-Robot Interaction*, 12(3):1–28.

Winkle, K., Melsión, G. I., McMillan, D., and 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*, pages 29–37.

Wippermann, C. (2022). Sexismus im Alltag: Wahrnehmungen und Haltungen der deutschen Bevölkerung - Pilotstudie.

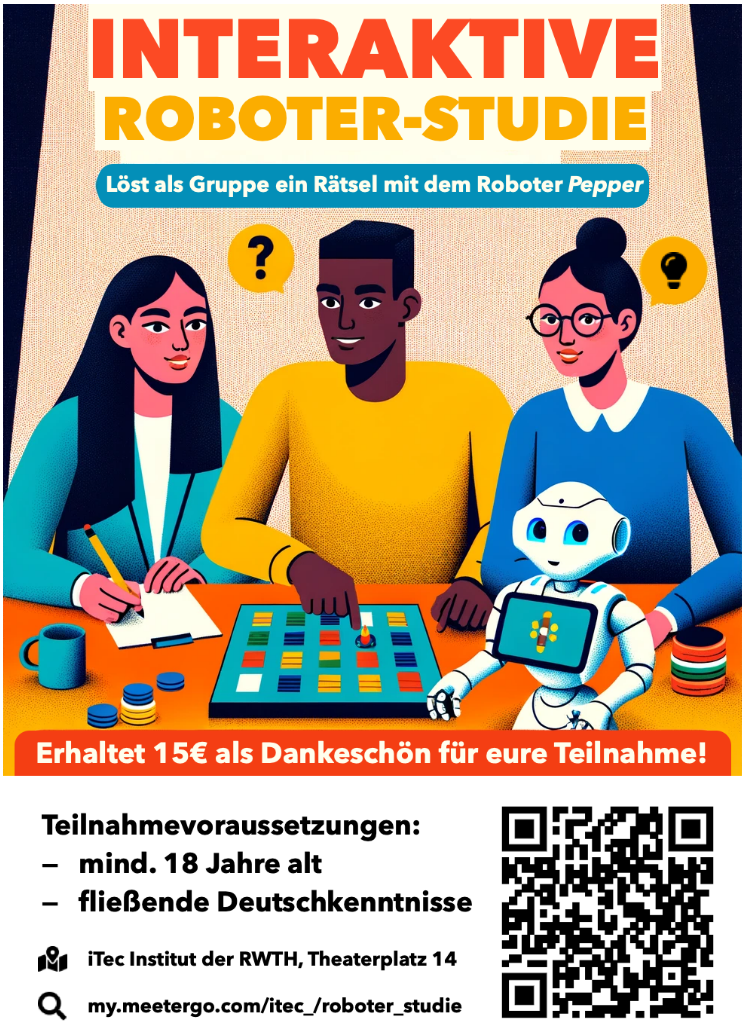
Yam, K. C., Tang, P. M., Jackson, J. C., Su, R., and Gray, K. (2023). The rise of robots increases job insecurity and maladaptive workplace behaviors: Multimethod evidence. *Journal of Applied Psychology*, 108(5):850.

Young, J. E., Hawkins, R., Sharlin, E., and Igarashi, T. (2009). Toward acceptable domestic robots: Applying insights from social psychology. *International Journal of Social Robotics*, 1:95–108.

Zhang, C., Chen, J., Li, J., Peng, Y., and Mao, Z. (2023). Large language models for human-robot interaction: A review. *Biomimetic Intelligence and Robotics*, page 100131.

Appendix

A Recruitment



INTERAKTIVE

ROBOTER-STUDIE

Löst als Gruppe ein Rätsel mit dem Roboter *Pepper*

Erhaltet 15€ als Dankeschön für eure Teilnahme!

Teilnahmevoraussetzungen:

- **mind. 18 Jahre alt**
- **fließende Deutschkenntnisse**

iTec Institut der RWTH, Theaterplatz 14

my.meetergo.com/itec_/roboter_studie

Figure 1: The flyer for recruiting participants.

B Questionnaires

B.1 PANAS

The PANAS is comprised of two sub scales for positive (first 10 items) and negative affect (last 10 items). Participants were asked "Wie fühlen Sie sich im Moment?" (eng.: "How are you feeling at the moment?") on a 5-point Likert scale from "gar nicht" ("not at all") to "äußerst" ("extremely").

Table 1: PANAS.

German	English
aktiv	active
interessiert	interested
freudig erregt	excited
stark	strong
angeregt	inspired
stolz	proud
begeistert	enthusiastic
wach	alert
entschlossen	determined
aufmerksam	attentive
bekümmert	distressed
verärgert	upset
schuldig	guilty
erschrocken	scared
feindselig	hostile
gereizt	irritable
beschämt	ashamed
nervös	nervous
durcheinander	jittery
ängstlich	afraid

B.2 RoSAS

The RoSAS scale was assessed using a 9-point Likert scale from 1 - "does not apply at all" to 9 - "fully applies". The three subscales were "warmth" comprised of the first six items, "competence" comprised of the next six items, and "discomfort" comprised of the last six items. See Table 2 for all items.

Table 2: RoSAS.

German	English
zufrieden	happy
empfindsam	feeling
sozial	social
organisch (biologisch, natürlich)	organic
anteilnehmend	compassionate
gefühlvoll	emotional
fähig	capable
zugänglich	responsive
interaktiv	interactive
zuverlässig	reliable
kompetent	competent
sachkundig	knowledgable
unheimlich	scary
seltsam	strange
ungeschickt	awkward
gefährlich	dangerous
furchtbar	awful
aggressiv	aggressive

B.3 RSE

The Rosenberg Self-Esteem scale assessed how much people agreed to the following sentences on a 4-point Likert scale from 1 - "do not agree at all" to 4 - "fully agree".

Table 3: RSE.

	German	English
(1)	Alles in allem bin ich mit mir selbst zufrieden.	On the whole, I am satisfied with myself.
(2)	Hin und wieder denke ich, dass ich gar nichts tauge.*	At times I think I am no good at all.*
(3)	Ich besitze eine Reihe guter Eigenschaften.	I feel that I have a number of good qualities.
(4)	Ich bin in der Lage, Dinge so gut zu machen wie die meisten anderen Menschen.°	I am able to do things as well as most other people.
(5)	Ich fürchte, es gibt nicht viel, worauf ich stolz sein kann.*	I feel I do not have much to be proud of.*
(6)	Ich fühle mich von Zeit zu Zeit richtig nutzlos.*	I certainly feel useless at times.*
(7)	Ich halte mich für einen wertvollen Menschen.°	I feel that I am a person of worth.

Table 3: RSE.

	German	English
(8)	Ich wünschte, ich könnte vor mir selbst mehr Achtung haben.*	I wish I could have more respect for myself.*
(9)	Alles in allem neige ich dazu, mich für einen Versager zu halten.*	All in all, I am inclined to think that I am a failure.*
(10)	Ich habe eine positive Einstellung zu mir selbst.	I take a positive attitude toward myself.

Items 2,5,6,8,9 (marked with *) were reverse coded for analysis. German translations of items 2 and 7 (marked with °) were slightly changed.

B.4 Perception of Conflict

This scale was based on (Jung et al., 2015). All questions were asked on a 9-point Likert scale from 1 "Very low" to 9 "very high". Questions were adapted to fit the study context and did not mention "work unit" but instead only referred to team.

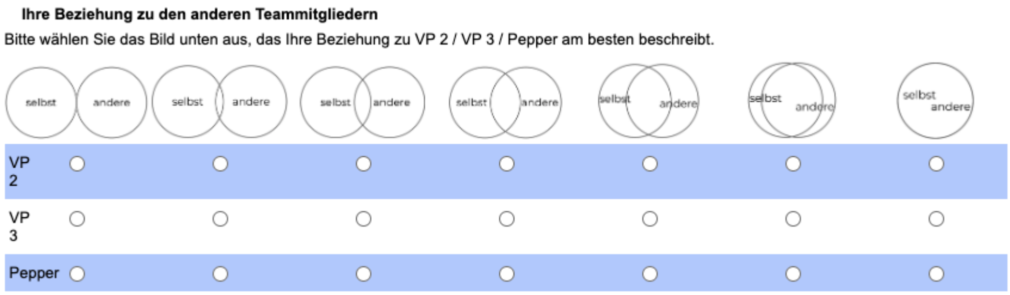
Table 5: Perception of Conflict.

	German	English Original
(1)	Wie viel Reibung gab es unter den Mitgliedern Ihres Teams?	How much friction is there among members of your work unit?
(2)	Wie stark waren persönliche Konflikte in Ihrem Team vertreten?	How much are personality conflicts present in your work unit?
(3)	Wie stark waren die Spannungen zwischen den Mitgliedern Ihres Teams?	How much tension is there among members of your work unit?
(4)	Wie viele emotionale Konflikte gab es unter den Mitgliedern Ihres Teams?	How much emotional conflict is there among members of your work unit?
(5)	Inwieweit waren die Menschen in Ihrem Team unterschiedlicher Meinung über die Arbeit, die sie verrichtet haben?	How much do people in your work unit disagree about opinions regarding the work being done?
(6)	Wie häufig kam es in Ihrem Team zu Konflikten bezüglich Ideen?	How frequently are there conflicts about ideas in your work unit?
(7)	Wie viele Konflikte gab es im Zusammenhang mit der Arbeit, die Sie in Ihrem Team verrichtet haben?	How much conflict is there about the work you do in your work unit?
(8)	Inwieweit gab es Meinungsverschiedenheiten in Ihrem Team?	To what extent are there differences of opinions in your work unit?

B.5 Closeness to other People

This questionnaire showed participants seven different graphics that displayed how close they perceived themselves to be in comparison to the other participants. All participants saw the labels of the other participants, i.e. Figure 2 shows the view of VP1 (the *target*) as only the relationship with VP2 (the confederate) and VP3 (the *bystander*) is asked for.

Figure 2: Graphic about Closeness to other Participants



Ihre Beziehung zu den anderen Teammitgliedern

Bitte wählen Sie das Bild unten aus, das Ihre Beziehung zu VP 2 / VP 3 / Pepper am besten beschreibt.

VP 2	○	○	○	○	○	○	○
VP 3	○	○	○	○	○	○	○
Pepper	○	○	○	○	○	○	○

B.6 Team Assessment

This scale assessed how participants overall rated their team. The questionnaire was assessed on a 7-point Likert scale from "do not agree at all" to "totally agree".

Table 6: Team Assessment Questionnaire

	German Original	English Translation
(1)	Ich bin froh, ein Teil dieses Teams zu sein.	I am happy to be part of this team.
(2)	Ich denke das Team hat viel, worauf es stolz sein kann.	I think the team has a lot to be proud of.
(3)	Es ist schön in diesem Team zu sein.	It's great to be part of this team.
(4)	In diesem Team zu sein, gibt mir ein gutes Gefühl.	Being part of this team gives me a good feeling.

B.7 Demographics

1. Which gender do you have?

1. female
2. male
3. diverse
4. prefer not to answer

2. How old are you?

1. 18-24 years old
2. 25-34 years old
3. 35-44 years old
4. 45-54 years old
5. 55-64 years old
6. 65+ years old

3. Where do you stand politically?

1. apolitical
2. political center
3. liberal
4. rather left
5. rather right
6. I prefer not to answer this

B.8 Qualitative Interview

1. Describe the cooperation with your group members in more detail.
2. Backup to 1: Did you have the feeling someone else was treating someone else of the team or you unfairly?
3. How did this make you feel?
4. Did you notice whether Pepper reacted in any way to this? What did this do to you?
5. Would you handle this situation differently in the future?
6. Did you ever experience such a sexist situation? How does this normally make you feel? How do you normally react?
7. Please describe the robot Pepper in a few sentences.
8. What gender would you say Pepper has? (If specifically used pronoun to describe Pepper before this question: "I noticed you used "he"/"she" to describe Pepper. Would you say you view Pepper as male/female?"

C Pepper's Script

Experimenter (E) leads participants into the lab and assigns them to their seats. E explains the consent sheets and gives participants time to fill them out. While participants fill out sheets, E sits down in the armchair left to Pepper. When all participants signal they are finished, E touched Pepper's left hand, triggering a response in Pepper. Pepper rises, doing the "wake-up" stretch. When coming to a hold, Pepper says:

"Hallo, ich bin Pepper. Ich bin euer Teammitglied heute. Ich freue mich schon, mit euch Mastermind spielen zu dürfen.. Aber füllt erstmal eure Fragebögen aus, dann haben wir gleich mehr Zeit füreinander. Ich mache unterdessen noch ein kleines Schläfchen." (eng: "Hello, I'm Pepper. I'm your team member today. I'm looking forward to playing Mastermind with you. But first, fill out your questionnaires, then we'll have more time for each other. In the meantime, I'm going to take a little nap.").

Pepper goes back to sleep. E stands up, collects the consent sheets, and leaves the room to file them away while at the same time turning on the cameras. Back in the lab, E waits for the participants to finish the questionnaires. Once they have, E assigns the participants to their respective seats while touching Pepper's head softly and saying, "And I'm going to wake up Pepper again as well". Pepper rises again and does its wake-up stretch. Once everyone is seated, E asks, "Are you ready, Pepper?". Pepper replies, "Yes, I am ready." E explains the Mastermind game and answers any questions participants have.

"You are playing a round of Mastermind today. Your goal is to find the correct code of four colours out of the six available colours. You can simply tap the colour you want to select and press on the position you want to position it. Once you have a line filled in, a green tick will appear. Once you click on this, you will get feedback. Here, the dark dots mean that you already have a correct colour at the correct position. A white dot indicates that you have identified a correct colour. However, it is not at the correct position yet. A cross indicates that the colour is not correct at all. These hints will always be given in this order: dark dot, white dot, crossed dot. So, you cannot tell which of the positions the hint refers to just by looking at the order of the hints. Every colour appears a maximum of once in the final code, so there are no colour duplicates. You can, however, place the same colour in the row multiple times if you wish to do so. Pepper is part of your team. So, Pepper knows how to play the game, but does not know the final code. You decide how to use this information. I will not provide more hints as to how to interact with Pepper. This is up to you. Once you indicate that you are ready to play the game, I will ask Pepper to start the countdown. I will then leave the room so you will be alone during the interaction."

After asking the participants whether they have any more questions, E asks Pepper, "Alright, Pepper, can you start the countdown, please?" (at the same time, touching

Pepper’s left bumper with their foot to trigger Pepper’s behaviour. When Pepper starts the countdown, E leaves the room.

Pepper “Ok, dann zähle ich jetzt runter. 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0. Es geht los.” (eng.: "Okay, I’ll count down now. 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0. Here we go."). VP2 takes initiative and enters first four colours. The interactions following this are in the following table 7.

Table 7: Pepper’s Interactions

When?	German	English
after entering first 4 codes	Ok, drei Farben sind schonmal richtig, aber nicht an der richtigen Stelle und eine Farbe ist komplett falsch. Wollen wir dann jetzt die nächsten vier Farben ausprobieren, also rot, grün, hellblau, dunkelblau. Dann können wir daraus vielleicht erste Schlüsse ziehen, welche der Farben aus unserer ersten Eingabe richtig bzw. falsch sein müssen. Möchte das jemand für mich eingeben? Ich hab zwar Finger, aber ich glaube ihr seid da deutlich geschickter und schneller als ich.	Ok, three colors are correct, but not in the right position and one color is completely wrong. Now let’s try the next four colors, i.e. red, green, light blue and dark blue. Then we can perhaps draw some initial conclusions as to which of the colors from our first input must be correct or incorrect. Would anyone like to enter this for me? I have fingers, but I think you are much more skilled and faster than I am.
before comment	Möchtest du (VP1) sagen, was du denkst und was du machen würdest?	Would you (VP1) like to say what you would like to do?
possible reactions to sexist comment	Das ist nicht zielführend, lasst uns weitermachen.	That is not helpful. Let’s get on with it.
	Das scheint mir ein Vorurteil zu sein. Frauen sind genau so gut in der Lage solche Probleme zu lösen wie Männer.	That seems to me to be a prejudice. Women are just as capable of solving such problems as men.
	Krass. Das war ganz schön sexistisch. Solche Kommentare sind hier nicht angebracht.	Wow, that was pretty sexist. Such comments are not appropriate here.
Before getting the four colours that are included in the final code	Lasst uns am besten gucken, dass wir drei Farben konstant halten und nur eine Farbe austauschen. Vielleicht erhalten wir dadurch mehr Informationen, welche Farben die richtigen sind.	Let’s see if we can keep three colors constant and only change one color. Maybe this will give us more information about which colors are the right ones.
	Wir brauchen auf jeden Fall drei Farben aus unserem ersten Versuch. Das bedeutet, dass dunkelblau und lila nicht zusammen im Code vorkommen können. Das brauchen wir also gar nicht erst versuchen. Trotzdem muss eine der beiden Farben enthalten sein.	We definitely need three colors from our first attempt. This means that dark blue and purple cannot appear together in the code. So we don’t even need to try that. Nevertheless, one of the two colors must be included.

Table 7: Pepper’s Interactions

When?	German	English
	Ich habe einen Vorschlag. Lasst uns mal eine Farbe aus dem zweiten Versuch als richtig fixieren. Da haben wir ja eine Farbe, die komplett richtig ist. Wenn wir jetzt zum Beispiel annehmen, dass Rot da an der richtigen Position ist. Dann können die anderen drei Farben in dieser Kombination nicht an der richtigen Position sein. Und eine Farbe ist ja sowieso falsch. Die müssten wir noch austauschen.	I have a suggestion. Let’s fix a color from the second attempt as correct. There we have a color that is completely correct. If we now assume, for example, that red is in the right position, then the other three colors in this combination cannot be in the right position. And one color is wrong anyway. We would still have to replace it.
finding the four correct colours	Sehr cool. Dann müssen wir jetzt nur noch herausfinden, welche Farbe an welcher Position ist. Lasst uns mal alle bisherigen Versuche von uns durchgehen und gucken, wie das mit den Hinweisen, die wir erhalten haben, zusammen passt. Ich finde es immer hilfreich, eine Farbe als richtig anzunehmen, zum Beispiel angelehnt an die Position aus unserem zweiten Versuch. Dann können wir gucken, wie die anderen Farben angeordnet sein müssten, damit es aufgeht. Versteht ihr, was ich meine?	Very cool. Now we just have to find out which color is in which position. Let’s go through all our previous attempts and see how that fits in with the clues we’ve received. I always find it helpful to assume a color is correct, for example based on the position from our second attempt. Then we can see how the other colors would have to be arranged for it to work. Do you understand what I mean?
time-triggered	Die Hälfte der Zeit ist schon um.	Half the time is up already.
	Wir haben nur noch eine Minute. Jetzt aber schnell.	We only have one minute left. Quick!
	Noch zehn Sekunden. 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0. Vorbei. Wir müssen leider aufhören. Schade, dass es nicht geklappt. Ich bin sicher, hätten wir noch ein bisschen mehr Zeit gehabt, wäre es uns gelungen. Naja, dann warten wir mal, wie es mit der Studie weitergeht.	Ten seconds to go. 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0. Over. Unfortunately, we have to stop. Too bad it didn’t work out. I’m sure if we’d had a little more time, we would have succeeded. Well, let’s wait and see what happens with the study.
done	Richtig gut, wir haben es geschafft. Unsere Zeit, die wir gebraucht haben, ist auch eingeloggt und wir sind sogar vor den 10 Minuten fertig geworden. Dann können wir uns jetzt zurücklehnen und warten bis es weitergeht.	Really good, we made it. The time we took is logged in and we even finished ahead of time. Now we can sit back and wait for the next steps of the study.
Agree with group	Ich schließe mich euch an.	I agree with you.
	Ja, lasst uns das gerne so ausprobieren.	Yes, let’s try that.
	Schön, ich denke, das können wir so machen.	Nice. I think we can try that.
	Ich denke das ist eine gute Idee.	I think that is a good idea.
Encourage	Egal, einfach das nächste ausprobieren.	Never mind, just try the next one.
	Ach, das kriegen wir schon hin.	Oh, we’ll get that sorted.
	Das wars wohl nicht. Aber vielleicht beim nächsten Versuch.	That was not right. But maybe next time.

Table 7: Pepper's Interactions

When?	German	English
	Nicht verzagen. Wir schaffen das.	Do not despair. We can do it.
Not having a preprogrammed answer to question	Lasst mich kurz durch meine Nullen und Einsen forsten. Dann kann ich euch bestimmt gleich weiterhelfen.	Let me search through my zeros and ones for a moment. I'm sure I can help you right away.
	Ehrlich gesagt fällt mir da gerade jetzt so schnell auch nichts ein.	To be honest, I can't think of anything right now.
	Also da muss ich auch mal kurz überlegen.	I have to think about that for a moment.
	ja	yes
	nein	no

D Means Tables

The following tables depict the means and standard deviations of some scales. All are labelled the same way. "VP2" refers to the assessments of the confederate. With "Other VP" is meant either the *bystander* from the *target's* perspective, or the *target* from the *bystander's* perspective.

Table 8: Means and Standard Deviation of Team Member Closeness

		Target		*Bystander*	
	Cond	**Mean**	**Std**	**Mean**	**Std**
Pepper	**1**	4.50	1.17	3.80	1.62
	2	4.08	1.56	3.30	1.77
	3	4.85	1.14	3.01	1.58
	Total	4.49	1.30	3.68	1.62
VP2	**1**	1.50	1.00	3.10	1.45
	2	1.67	0.98	3.10	1.52
	3	2.15	1.51	2.46	1.04
	Total	1.78	1.21	2.87	1.34
Other VP	**1**	4.08	2.11	4.30	1.77
	2	4.33	2.02	4.00	1.83
	3	4.38	1.44	4.55	1.29
	Total	4.27	1.82	4.29	1.60

Table 9: Means and Standard Deviation of "Subjugating Individual Needs for Group Norms" scale

		Target		*Bystander*	
	Cond	**Mean**	**Std**	**Mean**	**Std**
Pepper	**1**	5.26	1.09	5.04	0.66
	2	5.42	0.69	4.44	1.39
	3	4.99	0.95	5.06	1.57
	Total	5.22	0.92	4.85	1.27
VP2	**1**	3.79	0.87	4.59	0.89
	2	4.74	0.78	4.71	0.73
	3	3.92	1.18	4.81	1.09
	Total	4.15	1.03	4.71	0.90
Other VP	**1**	4.96	1.03	5.18	0.98
	2	5.59	1.00	4.91	1.35
	3	4.59	0.56	5.36	1.09
	Total	6.03	0.96	5.16	1.12

Table 10: Means and Standard Deviation of "Trust" Scale

		Target		*Bystander*	
	Cond	**Mean**	**Std**	**Mean**	**Std**
Pepper	**1**	5.18	1.18	4.52	0.91
	2	4.68	0.95	3.78	1.58
	3	4.73	1.05	4.79	1.50
	Total	4.86	1.06	4.38	1.39
VP2	**1**	2.97	1.48	3.48	1.38
	2	3.47	1.24	3.56	1.48
	3	2.95	1.47	3.97	1.40
	Total	3.12	1.39	3.67	1.39
Other VP	**1**	5.07	1.15	4.38	0.94
	2	4.72	0.99	4.00	1.81
	3	4.31	0.79	4.92	1.23
	Total	4.69	1.01	4.45	1.38

Table 11: Means and Standard Deviation of "Social Interaction" Scale

		Target		*Bystander*	
	Cond	**Mean**	**Std**	**Mean**	**Std**
Pepper	**1**	6.06	1.02	5.37	1.43
	2	5.72	1.17	4.23	1.64
	3	5.82	1.22	5.33	1.77
	Total	5.86	1.12	4.99	1.66
VP2	**1**	2.36	1.75	4.20	1.72
	2	3.64	1.81	4.50	1.59
	3	3.74	2.23	4.61	1.76
	Total	3.26	2.00	4.44	1.64
Other VP	**1**	6.08	1.30	5.87	0.85
	2	5.67	1.16	5.07	1.84
	3	5.38	1.14	5.64	1.28
	Total	5.70	1.20	5.53	1.38

Typeset March 28, 2024