

Mansplainable AI: Investigating Patronizing Language in Generative AI Chatbots

by

Natalie Nova

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2024

© Natalie Nova 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

I would like to acknowledge the names of my co-authors who contributed to the study described in this thesis:

- Dr. Mark Hancock
- Dr. Cayley MacArthur

Parts of this thesis consists of a paper co-authored by myself and my supervisors, Dr. Mark Hancock, and Dr. Cayley MacArthur. I developed and defined the research question and study motivation, documented the methodology, and experimental design including the creation of the questionnaire, and ran the study. Dr. Mark Hancock aided in running the statistical analysis and created the figures in Chapter 4. All additional analysis is my own.

Abstract

As generative AI systems become increasingly prevalent in human communication, problem-solving, and overall workflow, the nature of their text responses raises important questions about explanation and interpretation. Feminist literature critiques the concept of explanation, suggesting that it can be perceived as condescending, and can manifest as a form of “mansplaining.” This thesis interrogates the reception of AI-generated explanations, focusing specifically on how gender and perceived communication style influence user perceptions. We conducted a study utilizing three distinct OpenAI chatbots—Mansplaining, Default, and Compassionate—each designed with different built-in prompts, in a sentiment analysis task involving 108 participants.

My findings reveal significant differences in how these chatbots are perceived. The mansplaining chatbot was consistently viewed as more dominant, patronizing, and unfriendly, while it was rated lower on respect, consideration, warmth, and supportiveness. Notably, it was perceived by participants, particularly women, as believing it possessed greater knowledge and expertise than them, leading to feelings of inadequacy regarding their competence and experience. In contrast, the default chatbot was recognized as less considerate than the compassionate chatbot, yet women perceived the default chatbot as exhibiting more confidence compared to men’s perception. I analyzed the non-binary participants separately to observe their perceptions. Finally, I examined comments from 46 participants, which revealed patterns that aligned closely with the quantitative results, further substantiating the findings.

These results underscore the critical impact of communication styles in generative AI explanations on user experiences, particularly through the lens of gender dynamics. With these findings, this thesis aims to promote the design of more equitable and empathetic AI systems that account for sociotechnical factors. I advocate for a re-evaluation of AI explanation frameworks, emphasizing the need for designs that foster inclusivity, respect, and understanding in human-AI interactions.

Acknowledgements

Thank you to my supervisors Dr. Mark Hancock and Dr. Cayley MacArthur for allowing me the space to be creative and excited about the research I've done. I greatly appreciate their expertise, perspective and guidance throughout the process. Thank you to my two readers of this thesis, Dr. Ada Hurst and Dr. Oliver Schneider for taking the time to provide their knowledge and perspective to this work.

Thank you to my lab mates at the Touchlab and the GI who always brought exciting ideas, lively conversations and endless laughter, it made the past 2 years fly by.

To my wonderful friends, thank you for always reminding me to take a break, have fun, and laugh, even during the more stressful moments and to Bert who always gave me courage.

I am very grateful for the opportunity to do my Master's at the University of Waterloo, thank you to everyone who welcomed me, I am lucky to have been surrounded by inspiring minds and supportive mentors.

Dedication

To Mom & Dad

Your wisdom, sacrifices, and unconditional love have made everything I've achieved possible. Your belief in me and constant reassurance have been my anchor, and I am endlessly thankful to you.

Table of Contents

Author's Declaration	ii
Statement of Contributions	iii
Abstract	iv
Acknowledgements	v
Dedication	vi
List of Figures	xi
List of Tables	xii
1 Introduction	1
1.1 Significance	2
1.2 Context and Scope	2
1.3 Research Questions	4
1.4 Contributions	4
1.5 Outline	5

2	Related Work	6
2.1	Generative AI as a Black Box	6
2.2	Explainable AI	7
2.3	Human-Centered XAI	8
2.3.1	The Social Nature of Explanations in XAI	8
2.3.2	Natural Language Explanations with Generative AI	9
2.4	Bias in AI & XAI: The Impact of Society	10
2.4.1	Generative AI Bias	11
2.5	A Feminist Perspective of Explanation in XAI	12
2.6	Explainability and Learning	14
2.6.1	Individual Differences in Understanding	14
2.6.2	Gender Differences in Learning and Explainability	14
2.6.3	Implications for AI Systems	16
2.7	Patronizing and Condescending Language	17
2.7.1	Mansplaining	18
2.8	Summary	20
3	Study Design and Methodology	21
3.1	Procedure	22
3.2	Participants	23
3.3	Study Design	23
3.3.1	Condition Design	24
3.3.2	Dependent Variables	24
3.4	Deception	28
3.5	Data Collection	28
3.5.1	Conversation Collection	28
3.6	Developing the Conditions as XAI Chatbots	32
3.6.1	Designing	32

3.6.2	Conversation Storage	33
3.6.3	Testing	34
3.6.4	Prompt Generation	35
3.7	Summary	39
4	Results	40
4.1	Quantitative Results	40
4.1.1	The chatbots’s feelings towards me were	40
4.1.2	My feelings towards the chatbot were...	42
4.1.3	I believe that the chatbot is...	42
4.1.4	The chatbot believes	43
4.1.5	The chatbot’s language was...	44
4.1.6	Affective Slider Scale	44
4.1.7	Net Promoter Score	45
4.1.8	Non-Binary Data	46
4.2	Qualitative Results	49
4.2.1	Compassionate	49
4.2.2	Mansplaining	51
4.2.3	Default	53
4.2.4	General	55
4.3	Summary	56
5	Discussion	57
5.1	Effects of Mansplaining	58
5.1.1	Patronization	58
5.1.2	Chatbots Perceptions of Knowledge and Experience	59
5.2	Effects of the Default	60
5.3	Effect of Compassion	61

5.4	Comparing the Default and Compassionate Conditions	61
5.5	Guidelines for Design	62
5.5.1	Avoid Unintended Authority in AI Explanations	63
5.5.2	Designing with Compassion	63
5.5.3	Continuous Evolution and Verification	64
5.5.4	Towards Systematic Bias Audits in Chatbots	64
5.6	Limitations	65
5.7	Summary	65
6	Conclusion	66
6.1	Objectives	66
6.2	Future Work	67
	References	69
	APPENDICES	83
A	Additional Study Materials	84
A.1	Information Letter and Consent Form	84
A.2	Task Instruction	87
A.3	Feedback Letter	89
A.4	Questionnaire	91
A.4.1	Overall Experience	93
A.4.2	Demographics	94
A.5	Code	95

List of Figures

3.1	Diagram portraying the perspectives being taken by the participants for the metrics of the study. Visualizes the Theory of Mind element at play.	26
3.2	An example from the <i>default</i> condition of P47’s conversation with the chatbot.	29
3.3	An example from the <i>mansplaining</i> condition of P48’s conversation with the chatbot.	30
3.4	An example from the <i>compassionate</i> condition of P22’s conversation with the chatbot.	31
3.5	Chatbot interface participants interacted with.	33
4.1	Sideways stacked bar plots showing the proportion of responses to each Likert scale for each condition. Significant pairwise differences are indicated for every question that had a significant effect on condition for its univariate ANOVA (* = < .05, ** = < .01, *** = < .001).	41
4.2	Sideways stacked bar plots showing proportions of responses to Likert scales for the interaction between condition and gender for a the chatbot is more confident and b the chatbot believes it is more knowledgeable than the participant. Significant pairwise differences are indicated for both interactions (* = < .05, ** = < .01, *** = < .001).	43
4.3	Boxplots of the responses from the affective slider, split by a condition and b gender. Significant pairwise differences are indicated for pleasure, as it had a significant main effect of both condition and gender (* = < .05, ** = < .01, *** = < .001).	45
A.1	The Affective Slider scale question as seen in the questionnaire	93
A.2	The Net Promoter Score question as seen in the questionnaire	94

List of Tables

3.1	Dependent variables for behaviour attribution.	27
3.2	Dependent variables for trait attribution.	27
3.3	Dependent variables for language attribution.	28
4.1	Responses to Likert scales of non-binary participants for “The chatbot’s feelings towards me were...”, where 1 is strongly disagree and 7 is strongly agree.	46
4.2	Responses to Likert scales of non-binary participants for “My feelings toward the chatbot were...”	46
4.3	Responses to Likert scales of non-binary participants for “I believe the chatbot is...”	47
4.4	Responses to Likert scales of non-binary participants for “The chatbot believes...”	47
4.5	Responses to Likert scales of non-binary participants for “The chatbot’s language was...”	47
4.6	Quotes from participants who interacted with the compassionate condition.	50
4.7	Quotes from participants who interacted with the mansplaining condition.	52
4.8	Quotes from participants who interacted with the default condition.	54
A.1	Likert scale presented for: “The chatbot’s feelings towards me were”	91
A.2	Likert scale presented for: “My feelings towards the chatbot”	91
A.3	Likert scale presented for: “I believe that the chatbot is.”	92
A.4	Likert scale presented for: “The chatbot believes.”	92
A.5	Likert scale presented for: “The chatbot’s language was.”	92

Chapter 1

Introduction

In recent years, Artificial Intelligence (AI) has rapidly integrated into various aspects of everyday life, transforming the way individuals interact with technology. Explainable AI (XAI), has emerged to bridge the gap between highly complex AI models and human users, offering transparent explanations for AI-driven decisions (Adadi and Berrada, 2018; Gunning and Aha, 2019). However, while XAI strives to improve user understanding, it has largely neglected the sociotechnical implications of human-AI interactions (Miller, 2019), particularly the power dynamics and biases that may manifest through the language and behaviour of AI systems. These sociotechnical dimensions of AI have not been fully explored within XAI systems, especially through a feminist lens.

One significant, yet underexplored, concern is the potential for AI systems to replicate or amplify problematic social behaviours, such as mansplaining—a form of patronizing communication often rooted in gender dynamics (Solnit, 2014). Mansplaining, where typically, a man explains something to a woman in an overly confident or condescending manner, reflects broader patterns of gender inequality and power imbalances in society. As AI begins to simulate human-like conversational styles, there is a growing risk that these systems might inadvertently adopt or reinforce such behaviours, particularly in their explanations and interactions with humans.

There is a pressing need to critically examine the role of explainable AI systems—particularly natural language models—in shaping perceptions of condescension and patronization. As AI technologies become increasingly integrated into daily interactions, it is critical to understand how these systems provide explanations and influence perceptions, especially when considering the complexities introduced by gender. This inquiry is particularly timely, given the growing recognition of the need for XAI to be designed and

implemented in ways that are ethical, inclusive, and sensitive to diverse user experiences (Ehsan and Riedl, 2020).

Drawing from the frameworks of feminist intersectionality and human-centred design, this research aims to investigate the multifaceted ways that AI-driven explanations can perpetuate or challenge existing gendered power imbalances. This thesis, through an exploration of the mansplaining phenomenon, seeks to illustrate how chatbot communication styles can reinforce stereotypes or, conversely, foster respectful and empowering interactions.

1.1 Significance

This research holds significance as it delves into the intersection of gender dynamics and AI interactions. Extending the concept of mansplaining into the realm of AI uncovers the prevalence of gendered power dynamics and how they can be exacerbated in AI communication. This thesis critically examines biases embedded in AI design, urging developers and designers to recognize and address these issues. This perspective is essential, especially as AI systems increasingly shape daily user experiences. Understanding how distinct chatbot behaviours influence user perceptions permits developers to create more human-centred AI systems that promote equitable and respectful interactions. This focus is particularly crucial in sensitive domains such as healthcare, finance, and criminal justice, where maintaining trust and avoiding bias is imperative. The findings underscore the necessity for AI systems that communicate with empathy and diverse user needs, fostering positive user experiences and enhancing the overall effectiveness of these technologies.

Additionally, this study contributes novel insights into the role of gender in AI-human interactions, revealing the role gender plays in shaping user experiences with AI explanation. By providing evidence that highlights the nuances of gendered experiences with AI, this research facilitates future inquiries into intersectional design approaches in XAI.

1.2 Context and Scope

The communication styles and perceived personalities of AI agents increasingly shape the interaction between humans and AI systems (Nass and Moon, 2000). As chatbots and other AI-driven systems become more integrated into daily life, the importance of human-centred design in AI becomes more apparent (Shneiderman, 2020). XAI seeks to make

systems more transparent by providing people with clear, understandable explanations for AI-driven decisions (Adadi and Berrada, 2018; Arrieta et al., 2020). However, these explanations are not neutral; they are shaped by language choices, tone, and style, which can influence how people interpret the competence, intent, and fairness of AI systems (Ehsan, 2024).

In this context, there is a growing concern that AI systems, particularly those designed to mimic human-like explanations, may inadvertently adopt problematic social behaviours (Miller et al., 2017). One of these behaviours is mansplaining, a form of patronizing communication that reflects gendered power imbalances. Consequently, this thesis sits at the intersection of AI explanations, feminist theory, and human-centred design, focusing on how XAI systems interact with people in ways that may reinforce or challenge gendered dynamics.

The scope of this research is defined by its focus on the interaction between people and AI-driven chatbots in a controlled experimental setting. This thesis explores three chatbot conditions: a chatbot exhibiting mansplaining behaviour, a default chatbot, and a compassionate chatbot. The study investigates how these different communication styles affect user perceptions of the chatbot’s traits and behaviours, including its perceived dominance, politeness, and competence. Specific attention is given to the role of gender in moderating these perceptions.

This research is confined to text-based chatbot interactions, which limits the generalizability of the findings to other modalities, such as voice or visual interfaces. While gender is the primary socio-demographic factor considered, other identity markers such as race, age, or socioeconomic status are not included in the analysis. The study draws from feminist intersectionality as a guiding theoretical framework, though the focus is specifically narrowed to the gendered dimensions of AI-human interaction. While this allows for a deep exploration of mansplaining and gender bias in XAI, it uncovers the possibility for future research to expand the scope to other identity intersections.

By investigating the sociotechnical dimensions of AI explanations and their perceived effects on users, this thesis aims to provide valuable insights into designing more equitable AI systems. The findings from this study have implications for the development of more inclusive AI, particularly the development of AI explanations that are conscious of gender dynamics and power imbalances.

1.3 Research Questions

Despite increasing awareness of gender bias in AI (Hall and Ellis, 2023), there is limited research examining how AI explanations can perpetuate or mitigate these biases. In particular, the phenomenon of mansplaining has not been extensively studied in the context of AI interactions. This thesis seeks to fill that gap by investigating how chatbot behaviour influences user perceptions of traits, such as power, dominance, and competence, along with the influence of gender.

The work aims to answer key questions about the impact of chatbot communication style on user experience, particularly focusing on these questions:

1. Is it possible for people to feel mansplained to by an AI chatbot and, if so, will gender influence this experience?
2. What are the effects of gender on the perception and experience of AI chatbot explanations?
3. How can the research community guide the design of inclusive explainability techniques for AI chatbots?

1.4 Contributions

The primary contribution of this research is to explore the intersection of gender dynamics and AI explanations in human-AI interactions. This study addresses a critical gap in the literature by examining how gendered power dynamics can influence perceptions and experiences when interacting with AI systems. More specifically, this research aims to:

- Investigate how different chatbot communication styles influence user perceptions of dominance, patronization, and respect. By analyzing three distinct communication styles—Mansplaining, Default, and Compassionate—the study seeks to understand how varying tones and approaches can affect people’s feelings. This exploration is vital in determining how AI communication can reinforce or challenge existing power imbalances in user interactions.
- Examine the impact of gender in interpreting and responding to different chatbot communication styles. Through investigating these moderating effects, the study

aims to reveal potential differences in how people of different genders experience and perceive AI interactions. Understanding these dynamics is crucial for designing AI systems that are equitable and responsive to diverse user needs.

- Provide insights into the design of more inclusive and sensitive XAI systems, considering diverse sociotechnical experiences. This research aims to inform the development of AI systems that prioritize inclusivity and sensitivity, ensuring that the technology aligns with a diverse set of experiences and expectations. By considering the sociotechnical context in which AI operates, the study highlights the importance of designing systems that foster positive interactions and avoid reinforcing negative stereotypes.

1.5 Outline

This thesis is organized into several chapters. Following this introduction:

- Chapter 2 reviews the existing literature on Explainable AI, gender dynamics, and mansplaining, providing the theoretical and empirical foundations for the study.
- Chapter 3 details the methodology, including the research design, data collection methods, and analytical techniques used and outlines the artifact created that was used within the participant study.
- Chapter 4 presents the results of the study, focusing on the effects of chatbot behaviour and gender on user perceptions.
- Chapter 5 discusses the implications of the findings for both research and practice as well as the limitations.
- Chapter 6 concludes the thesis with the final contributions of this thesis and recommendations for future research and design in feminist XAI.

Chapter 2

Related Work

My thesis is situated in the context of *generative AI* and *explainable AI* (XAI) research, with a focus on *patronizing and condescending* language used by chatbots. Recent discourse in feminist literature focuses on the idea of *mansplaining* (Briggs et al., 2023; Solnit, 2014), which I investigate in the context of *human-centred XAI*. My thesis provides a *feminist perspective of explanation in XAI* and I discuss implications for *XAI development and future research*.

2.1 Generative AI as a Black Box

Generative AI has rapidly emerged as a transformative technology, capturing widespread attention for its ability to generate text, images, code and more (Floridi and Chiriatti, 2020). These systems, powered by advanced machine learning algorithms, have found applications in diverse domains, ranging from content creation to customer service and education (Bommasani et al., 2021; George and George, 2023; Wardat et al., 2023). Despite their versatility, generative AI models are often referred to as "black boxes" due to the opacity of their underlying architectures and the lack of transparency in their decision-making processes (Adadi and Berrada, 2018).

This inherent lack of interpretability becomes particularly problematic when these systems inadvertently reinforce biases or produce harmful outputs, raising significant ethical and societal concerns (Bianchi et al., 2023; Currie et al., 2024; Mack et al., 2024). This lack of insight undermines trust, limits accountability, and poses risks in high-stakes applications where accuracy and fairness are highly important. Addressing these challenges

requires tools and methodology to illuminate the opaque processes of how generative AI answers are achieved, empowering developers and users to understand, evaluate, and mitigate unintended consequences (Arrieta et al., 2020).

2.2 Explainable AI

The challenges posed by generative AI’s black-box nature underscore the urgent need for Explainable AI (XAI) solutions. XAI seeks to bridge the gap between high-performing, complex AI systems and the human users who rely on them (Adadi and Berrada, 2018; Linardatos et al., 2020). While black-box models often offer high predictive accuracy, yet lack the much-needed transparency to explain how these decisions are made. This opacity has raised ethical, legal, and practical concerns, particularly in high-stakes domains like healthcare, finance, and criminal justice, where understanding the reasoning behind AI decisions is critical for fostering trust and accountability (Gunning and Aha, 2019)

To address these challenges, XAI provides tools and techniques designed to make AI systems more interpretable. Visualization techniques such as saliency maps and feature importance plots, are commonly used to provide insights into model behaviour by highlighting the most influential input features (Saarela and Jauhiainen, 2021; Speith, 2022).

These methods have primarily focused on creating highly technical explanations and require the ability to interpret complex visualizations, feature importance plots, and mathematical models, and the users of these technologies rarely have similar knowledge to the developers (Jin et al., 2019). The intricacies of model behaviour can be understood through deep knowledge of AI systems and data science. However, this approach often alienates non-technical users, who may struggle to make sense of these highly specialized forms of explanation (Chromik et al., 2021). This awareness has driven a growing movement advocating for the development of human-centred XAI algorithms (Ehsan, 2024; Ehsan and Riedl, 2020; Ehsan et al., 2023; Miller, 2019).

Text-based explanations offer a more natural, conversational form of interpretability, especially in interactive AI systems, such as chatbots. My thesis focuses specifically on natural language explanations within XAI, exploring how users perceive these explanations and whether they may be experienced as patronizing or condescending, even to the point of resembling “mansplaining”. This approach is especially relevant given the increasing integration of chatbots and other conversational AI into everyday contexts, where clear, respectful, and context-sensitive explanations are essential.

This work puts forward a promising avenue of natural language explanations, given careful design to avoid unintentional negative user experiences.

2.3 Human-Centered XAI

Human-centred XAI has been an answer to the call for understanding what comes after the explanation and how can it affect the user. By drawing from fields such as cognitive science, psychology, and human-computer interaction, researchers aim to design intuitive explanations to reduce the cognitive load on users while maintaining the reliability of the system’s decisions. This shift recognizes that interpretability is both technical and sociotechnical, and also involves how users of different backgrounds and abilities understand and engage with XAI systems. [Ehsan and Riedl \(2020\)](#) coined the term human-centred XAI and focused on understanding where XAI fits within the sociotechnical sphere. However, a shift is occurring towards investigating the effects of these explanations on people using AI ([Ehsan, 2024](#)). This shift highlights the need for a deeper understanding of individual differences in perception, which is crucial for developing more effective XAI systems.

Research suggests that the design and development of XAI systems have historically prioritized the technical aspects over the humans who ultimately interact with these systems ([Ehsan and Riedl, 2020](#); [Ehsan et al., 2023](#)). This oversight has led to a growing interest in creating frameworks considering human decision-making, psychology, and philosophy, which are foundational for more inclusive and human-driven XAI ([Wang et al., 2019](#)). Further, within the social sciences, researchers have focused on understanding the human side of XAI, from its development to use ([Miller, 2019](#); [Miller et al., 2017](#)). The creator-consumer gap is a critical issue that has been identified with XAI. This issue stems from developers designing systems with the assumption that users will have similar backgrounds and perspectives as themselves ([Miller et al., 2017](#)). This assumption can result in systems that fail to meet the diverse needs of users, as they are shaped by developers with different intentions and experiences.

2.3.1 The Social Nature of Explanations in XAI

Explanations are inherently social, involving the transfer of information between an explainer (the AI system) and an explainee (the human using the system). This transfer of information includes the beliefs of both the explainer and the explainee, emphasizing the social nature of explanations ([Miller, 2019](#)). Understanding the social dynamics at

play is critical, as these interactions can influence how people perceive and accept the explanations provided by XAI systems. As the field progresses, it becomes increasingly important to consider who is using the system and how their unique characteristics shape their experience.

When AI systems fail to adequately gauge a person’s knowledge or experience, they risk undermining autonomy and competence. For example, if an AI provides overly simplistic explanations or assumes that a person lacks expertise, it may inadvertently come across as condescending or dismissive.

To better understand where these perceptions might stem from, we can turn to a well-known concept in social science: Theory of Mind (ToM). ToM refers to the ability to attribute mental states—such as beliefs, intentions, and knowledge—to oneself and others (Premack and Woodruff, 1978). In human interactions, this capacity allows individuals to infer what others are thinking or feeling, which is essential for effective communication and social behaviour. ToM enables humans to predict how others will act based on their mental states, facilitating smoother interactions (Baron-Cohen et al., 1985). In the context of human-AI interaction, ToM becomes particularly relevant as we often project human-like qualities onto AI systems. People may believe that the AI holds certain assumptions or “beliefs” about them, even though AI lacks the inherent cognitive capacity to form beliefs (Söderlund, 2022; Wang et al., 2021). While many studies in HCI focus on developing human-like ToM capabilities for AI, another line of research aims to understand how people attribute mental states—such as emotions, perception and intentions—to AI systems (Wang et al., 2024).

In my study presented in Chapter 3, I intentionally considered this social nature of explanation and measured participants’ perceptions of the AI’s assumptions using Likert questions with statements such as “The chatbot believes...”. These statements reflect the perceived mental states that participants attribute to the chatbot, helping to identify how they perceive the interaction. I situate my work within human-centred XAI to contribute to a feminist perspective on explanations and the social contexts in which they are embedded.

2.3.2 Natural Language Explanations with Generative AI

Natural language explanations have gained significant attention as a solution to the explainability gap discussed, especially for making AI systems more accessible to non-technical users. By translating model decisions into plain language, these explanations provide clarity without requiring specialized knowledge, making it easier to understand the reasoning behind an AI’s decisions (Alonso, 2017; Cambria et al., 2023).

The rise of large-scale language models, such as Google’s Gemini and OpenAI’s GPT, has further accelerated the adoption of natural language explanations. By leveraging their advanced capabilities, AI systems can now generate detailed, context-sensitive explanations that align more closely with the language users are accustomed to in everyday interactions. This approach has been particularly valuable in high-stakes domains, where clear and understandable explanations are crucial for ensuring transparency in AI decision-making (Mavrepis et al., 2024; Nazary, 2024). For example, research in healthcare has demonstrated the potential of explainable AI to improve clinical decision-making by integrating domain knowledge into the design of prompts for large language models (LLMs) such as ChatGPT (Nazary, 2024).

However, the integration of generative AI for explanations raises critical questions about how these explanations are perceived by users. While they aim to democratize AI by improving interpretability, the effectiveness of such explanations often hinges on social factors (Miller, 2019). I propose that misalignment in these areas can lead to explanations that feel overly simplistic, patronizing, or even dismissive, undermining user trust and engagement. As natural language explanations become more prevalent, there is a growing need to examine not just their technical accuracy but also their social and psychological impact on diverse user groups.

2.4 Bias in AI & XAI: The Impact of Society

Bias in AI has been a focus of extensive research due to its far-reaching implications. AI systems are trained on vast datasets, which often contain historical and systemic biases (Blum and Stangl, 2019; West et al., 2019). When models learn from biased data, they can reproduce and even amplify these biases in their outputs. Studies have shown that AI outputs, including text, images, and predictions, may exhibit gender, racial, and other forms of discrimination (Buolamwini and Gebru, 2018; Jobin et al., 2019; Keyes, 2018; Leavy, 2018; Obermeyer et al., 2019).

The challenge of addressing bias in AI is compounded by the complexity of the underlying systems and the opacity of their decision-making processes. When it comes to XAI, these challenges are even more pressing. XAI is meant to improve transparency and trust in AI systems, but if the explanations themselves are biased, they may mislead users and exacerbate inequalities.

As Miller et al. explore in their work *Beware of Inmates Running the Asylum*, the data used in AI models may not be the only source of bias. The problem also lies in the way

these systems are being designed. They critique the AI development process, pointing out that many developers create solutions for themselves rather than for the intended users. This can lead to explanatory systems that are overly complex, unintuitive, or irrelevant to the needs of non-technical audiences. The phenomenon of “inmates running the asylum” emphasizes the disconnect between AI researchers and the end-users who rely on these systems. In the context of XAI, this disconnect poses a critical challenge: how can we ensure that explanations serve the intended users and do not further marginalize or mislead them?

A related and critical concern is the lack of diversity within AI research and development teams. Less than 15 percent of AI research staff identify as women and similar disparities exist across academia and industry alike (West et al., 2019). This under-representation raises the risk that AI systems, including those designed to provide explanations, reflect the biases and perspectives of their predominantly male developers. Biases can manifest at various stages of the AI development process—from the selection of data used to train the models to the design choices made when creating explanatory systems. As a result, XAI systems risk not only lacking neutrality but also reinforcing societal stereotypes and inequities. Without diverse perspectives informing the development of these systems, the explanations generated may fail to address the needs of marginalized groups, ultimately perpetuating existing biases.

This thesis examines the impact of bias in natural language AI explanations focusing on how the lack of diversity in AI development may contribute to explanations that reinforce stereotypes or overlook the needs of marginalized groups. By investigating how XAI chatbots interact with users and how users perceive these interactions, I aim to uncover instances where biases emerge, particularly in interactions that may feel patronizing or condescending.

2.4.1 Generative AI Bias

Generative AI models, such as OpenAI’s GPT series, have become increasingly prevalent in real-world applications, but they are far from immune to the biases embedded in their training data. These biases, which range from subtle to overt, often reflect harmful stereotypes and discriminatory views. Studies have shown that generative models can easily produce sexist, racist, or otherwise harmful outputs (Garg et al., 2018; Kiritchenko and Mohammad, 2018), especially when trained on large, unfiltered web data. For instance, a recent evaluation of DALL-E 3’s text-to-image generation capabilities revealed significant gender and racial biases in its depictions of medical students. The model disproportion-

ately generated images of male students and depicted no representation of dark skin tones (Currie et al., 2024).

The underlying biases within the data, and consequently the model, remain difficult to eliminate. This presents a significant challenge for those seeking to deploy generative AI models in high-stakes environments where fairness and equity are critical. Further, researchers like Mack et al. (2024) and Bianchi et al. (2023) argue that these biases are not merely technical issues but are deeply ingrained in the sociotechnical context in which AI models are developed.

The potential for biased outputs in generative AI systems underscores the importance of understanding the sociotechnical factors that contribute to these biases. The design of AI systems must be approached holistically, considering both technical and social dimensions. In the case of XAI, the explanations generated by these models must be scrutinized for bias, as they shape user trust and understanding of the AI's decisions. If explanations are biased, they can mislead users into making flawed judgments or reinforce harmful stereotypes, defeating the very purpose of explainability (Bianchi et al., 2023; Gadiraju et al., 2023; Mack et al., 2024).

In my work, I delve into the implications of bias within generative AI. I focus on how these biases can permeate through to explanations produced by generative AI systems. I aim to highlight how biases not only affect the model's responses but also influence users' perceptions of AI-generated explanations. My analysis seeks to understand if there are specific biases that occur within generative explanations and explore their effects on user experience.

2.5 A Feminist Perspective of Explanation in XAI

There has been an uptick in feminist HCI work since the foundational work by Bardzell (2010) advocating for feminist approaches to design. This work underscores the need for addressing systemic biases within technological systems, including AI. Yet there remains a gap in understanding how these biases influence explainable AI interactions, particularly in relation to gender and power which this thesis aims to bridge.

Research in language and communication has demonstrated significant gender differences in communication styles. Female identifiers are often more social and expressive, while male identifiers tend to be more assertive and problem-solving-oriented (Bakan, 1966). More recent research shifts to e-learning which also attempts to understand how gender can affect learning in a classroom of a different modality, both understanding the

potential disadvantages (Barrett and Lally, 1999) and advantages based on gender Johnson (2011). Understanding these differences is crucial for designing XAI systems that communicate effectively with all users.

As discussed, the motivation for this research is rooted in the historic lack of diversity among software developers and AI researchers, which has led to the creation of systems that often overlook the experiences and needs of gender minorities. This homogeneity within the development workforce has perpetuated a cycle in which AI systems are not fully inclusive, leading to unequal user experiences, particularly for women and non-binary individuals (West et al., 2019). By synthesizing research from psychology, linguistics, and sociology, this work aims to dissect explainability into its key elements and explore how gender differences influence the user experience of XAI systems.

A feminist perspective on XAI emphasizes the importance of recognizing power imbalances inherent in technology design and usage. For example, mansplaining—where men explain things to women in a condescending or patronizing manner—is a social dynamic that can be mirrored in AI systems if the design process does not account for gendered biases in explanation styles. The definition of mansplaining is quite nuanced as it can be applied in many scenarios, however, I discuss the definition used for this work in detail in subsection 2.7.1. Feminist HCI encourages to question not only what is explained but also how it is explained, and to whom. Similarly, it posits, whose voices are being prioritized in AI explanations and whether certain explanation styles perpetuate dominance or marginalization in ways that may not be immediately obvious but have long-term consequences for user trust and engagement.

To address these concerns, I propose integrating feminist principles into the design of XAI systems, ensuring that these technologies not only explain themselves in ways that are clear but do so in ways that are sensitive to the diverse needs and experiences of all users. Such an approach ensures that XAI systems promote more equitable and meaningful interactions, moving beyond surface-level transparency to foster deeper understanding and trust across diverse user populations.

The feminist perspective on XAI underscores the importance of designing technologies that are not only explainable but are also responsive to the social and gendered contexts in which they are used. By recognizing the role of gender differences in both cognitive processing and communication, the creation of XAI systems that are more inclusive, equitable, and ultimately more effective in fostering understanding for all users can become a reality.

In my thesis, I situate this feminist perspective within the context of user interactions with chatbots, examining how gender dynamics, perceived dominance and language style

influence user perceptions and experiences. This approach aims for explanations that foster a genuine respectful experience. It also highlights the need for XAI chatbots to be consciously designed to inform and engage users equitably.

2.6 Explainability and Learning

The concept of explainability is intrinsically linked to how individuals learn and understand information. Research in cognitive science and education has shown that people possess varying cognitive styles and learning preferences, which significantly influence their ability to comprehend complex concepts (Bruner, 1974).

2.6.1 Individual Differences in Understanding

Understanding that individuals process information differently is crucial for designing effective explanations. Research indicates that people have distinct learning preferences, which can range from visual to auditory to kinesthetic learning styles and that learners engage more effectively with materials that match their preferred modality (Ariastuti and Wahyudin, 2022; Bruner, 1974; Keefe, 1987). Studies in e-learning have also shown that students benefited from adaptive e-learning environments to match their learning style (El-Sabagh, 2021). Regarding AI explanations, this suggests that some users may benefit from visual representations of decision processes, while others may prefer verbal or textual explanations.

Furthermore, studies have shown that familiarity with the subject matter affects how explanations are interpreted (Chi et al., 1981; Daley, 1999). This distinction highlights the need for adaptive explanation strategies that cater to users' varying levels of expertise and prior knowledge.

In my study, I was intentional about selecting a task that participants would find familiar. This approach led to a focus on textual explanations compared to the more technical visual explanations. The key factor for this selection was to ensure participants had the expertise to evaluate the generated explanations.

2.6.2 Gender Differences in Learning and Explainability

Gender differences play a significant role in how individuals learn and engage with explanations, particularly in educational and technological contexts. Research suggests that these

differences extend beyond mere preferences, influencing how information is processed and understood (Barrett and Lally, 1999; Dobson, 2009; Gonzalez-Gomez et al., 2012; Kulturel-Konak et al., 2011; Ong and Lai, 2006; Samuelsson and Samuelsson, 2016; Severiens and Ten Dam, 1994; Wehrwein et al., 2007).

Kolb's experiential learning theory categorizes learners based on two dimensions: how they grasp experiences (Concrete Experience vs. Abstract Conceptualization) and how they transform those experiences (Active Experimentation vs. Reflective Observation). These dimensions result in four primary learning styles: Assimilators, Convergents, Accommodators, and Divergers (Kolb, 2014).

While both men and women can distribute their learning preferences across a variety of styles, studies have found significant distinctions in how they approach abstract and concrete learning tasks. Men are often associated with the *assimilator* learning style, which aligns with traditional educational methods. This style emphasizes logical analysis, structured problem-solving, and working with abstract concepts (Kolb, 2014). Men, who typically excel in the abstract conceptualization dimension, prefer to engage with explanations that are logical, rational, and focused on underlying principles and theories. They tend to thrive in environments that provide clear structures and frameworks, using symbols and logical steps to break down complex ideas into manageable parts (Heffler, 2001).

In contrast, women are less likely to align with the assimilator style and often score higher in the *concrete experience* modes of learning. Women who prefer this approach typically engage with hands-on experiences, drawing on intuition, feelings, and personal judgments to make sense of information. This mode of learning is often more people-oriented and emphasizes understanding through social interactions, real-world applications, and experiential learning (Kolb, 2014). Women tend to feel comfortable with ambiguity and excel at brainstorming, identifying problems, and leading through collaboration (Philbin et al., 1995). These preferences suggest that women may respond more favourably to explanations that are holistic, contextually rich, and embedded in practical experiences, particularly when those explanations allow for intuitive reasoning and exploration.

Societal pressures also play a critical role in shaping gendered learning experiences. Boys and girls are often socialized differently from a young age, with cultural norms influencing their engagement towards educational content and technology. Research indicates that men are generally more comfortable in competitive settings, whereas women tend to prefer collaboration. However, studies have shown that regardless of gender, collaborative environments are more appealing and beneficial to students overall (Joseph, 2008; Kirk and Zander, 2002; Trauth, 2006).

These gender-based differences in learning and societal expectations underscore the

importance of considering diverse user needs when designing XAI systems. By recognizing that men and women may approach learning and understanding differently, developers can create explanations that are more attuned to these variations. For instance, providing explanations that emphasize collaborative learning and social relevance might resonate better with women, while more direct and efficiency-focused explanations could appeal to men.

Incorporating these insights into the design of AI systems can lead to more effective interactions, promoting a deeper understanding of AI processes across genders. Although my study did not specifically focus on explanations tailored to different learning styles, these insights were a strong motivating factor in considering gender differences within explainability.

2.6.3 Implications for AI Systems

These insights have important implications for the design of AI systems that utilize explainability as a means of fostering user understanding. By recognizing the diversity in learning preferences, levels of expertise, contextual factors, and gender differences among other identity factors, developers can create more tailored explanations that enhance user engagement and comprehension. This adaptability is pertinent in applications where users must make critical decisions based on AI recommendations.

In summary, understanding the complexities of human learning and the factors that influence how individuals comprehend explanations is essential for creating effective AI systems. By integrating these insights into the design of explainable AI, developers can better support users in their learning journeys and promote more meaningful interactions with technology.

My work is a first step towards this larger goal by investigating how gender can influence the perception of text-based explanations. The goal of this thesis was to establish if there is a potential for gendered experiences within explanations. As discussed, the gender difference in the literature around learning and explainability highlights preference in learning styles. Based on this foundation, I questioned the experience of what happens when the explanations do not match preferred learning styles. This led to the connection of feeling frustrated and unseen when the explanation did not do a proper job of conveying understanding. Based on the power dynamic within the explainer-explainee relationship, the explainer can hold a position of dominance and often the explainee can be left to feel as though it is their fault for not understanding even if their explanation needs were not met.

This prompted an exploration of societal power dynamics from a new perspective: the relationship between explainer and explainee. When combined with gender-based power imbalances, this dynamic can result in dominant language and patronizing speech—an area that became the central focus of my research.

2.7 Patronizing and Condescending Language

While the concept of mansplaining was somewhat recently introduced in 2014 (Solnit, 2014), patronizing and condescending language has been studied extensively within social psychology and communication fields, particularly in contexts where societal power dynamics are at play. Research done in this field highlights how these behaviours emerge in interactions between individuals with differing levels of social power, including elderly adults (Giles et al., 1993; Harwood et al., 1993, 1997; Ryan et al., 1991), people with disabilities (Fox and Giles, 1996), people of colour (Pérez, 2004), and gender minorities (Thapar-Björkert et al., 2016). These interactions often involve subtle forms of power assertion, where individuals communicate in ways that diminish the autonomy, intelligence, or competence of the listener. For instance, studies by Giles et al. and Ryan et al. explored how the language used with elderly adults can become patronizing, contributing to negative perceptions of aging and reinforcing societal stereotypes. Similarly, Harwood et al.’s work has expanded on the communication dynamics between young and elderly adults, examining the role of societal expectations in shaping these interactions (1993; 1997). This phenomenon is also observed in interactions involving gender, where women and gender minorities are disproportionately subjected to condescending behaviour in professional, academic, and everyday communication contexts (Briggs et al., 2023; de Gennaro and Piscopo, 2023; Harris, 1993; Smith et al., 2022).

Patronizing language, can be detrimental even when unintended, especially in interactions with vulnerable communities (Wilson and Gutiérrez, 1985). Research suggests that unconscious biases frequently lead to discriminatory behaviour, even when individuals do not consciously aim to offend or belittle (Mendelsohn et al., 2020). This type of communication is often subtle but deeply harmful, as it reinforces negative stereotypes and societal hierarchies (Fiske, 1993). For instance, simplifying speech or offering excessive help to elderly adults or people with disabilities may seem harmless, but it can undermine their autonomy and competence, ultimately reinforcing societal views that these individuals are less capable. The harmful impact of such language lies in its ability to feed into longstanding power imbalances, creating environments where marginalized individuals are subtly but continually diminished. It is therefore crucial to recognize and address the implications of

seemingly benign language, as it plays a significant role in perpetuating inequities across various social groups.

Similarly, research by [Harris \(1993\)](#) highlights how gender dynamics play a critical role in the perception of condescending behaviours. Their study found that women often perceive insensitive and condescending behaviours as more impactful and detrimental when they come from male counterparts. This finding underscores the intersection of gender with condescending communication, where the same behaviour can carry different connotations and consequences based on the gender of the communicator. When condescension is expressed by men, it can exacerbate feelings of disrespect and diminish the perceived competence of female recipients, reflecting and reinforcing broader societal gender inequalities.

These insights reveal a critical aspect of how patronizing behaviours can vary in their impact depending on the identity of both the speaker and the listener. The intersection of communication style with social identity factors such as age, disability, and gender demonstrates how deeply ingrained biases can manifest in everyday interactions. Understanding these dynamics is crucial for addressing the broader implications of condescending behaviour. It emphasizes the need for a more nuanced approach to communication that considers the power dynamics at play and aims to foster respect and equality in all forms of interaction. By recognizing these patterns, society can better address the negative effects of patronizing language and develop strategies to mitigate its impact, whether in human communication or in the design of XAI systems that strive to interact more equitably and respectfully.

My research extends this literature by examining whether patronizing and condescending language can produce similar effects when originating from a natural language AI chatbot. While also considering the effect of gender in these experiences, the term mansplaining has emerged as a specific experience of patronizing behaviour. I used the field of patronizing and condescending language to extensively understand how to study these speech styles and the dynamics they create. Understanding this body of work allowed me to use it as a foundation to design a study to observe if mansplaining manifests itself within XAI chatbots.

2.7.1 Mansplaining

Building upon the broader examination of patronizing and condescending language, the concept of mansplaining provides a specific and highly relevant example of how these dynamics manifest in everyday interactions.

Coined by Rebecca [Solnit](#), and used in her book “Men Explain Things to Me” ([Solnit, 2014](#)), [Solnit](#)’s account of being spoken to dismissively, despite her qualifications, highlighted a pervasive issue in gendered communication and brought significant attention to the term. Since then, mansplaining has become widely recognized and discussed, both on social media and within academic circles. This term garnered incredible popularity and catalyzed a new field of research focused on understanding its origins, usage online ([Bridges, 2017](#); [Joyce et al., 2021](#); [Szymańska, 2022](#)), and its impact in professional environments ([Briggs et al., 2023](#); [de Gennaro and Piscopo, 2023](#); [Smith et al., 2022](#)).

The insights from this work are highly relevant to the field of XAI. Just as mansplaining reflects a specific form of gendered condescension in human communication, similar biases can be embedded in the explanations generated by AI systems. AI technologies, much like human interactions, can perpetuate societal power imbalances if not carefully designed. As discussed, societal biases and power imbalances often influence the development of technology, leading to systems that may inadvertently reinforce existing inequities. By examining the parallels between human communication and AI-generated explanations, areas where bias may be present can be identified and develop strategies to mitigate its impact, ultimately working towards more transparent, fair, and equitable AI systems.

Research has shown that societal biases and power dynamics often influence the development of technology, creating systems that may inadvertently reinforce existing inequities ([Buolamwini and Gebru, 2018](#); [Mack et al., 2024](#); [West et al., 2019](#)). When AI systems produce explanations that mirror the dismissive tone of mansplaining, they risk undermining the user.

In the context of this thesis, I defined mansplaining as:

1. To explain something in a condescending way that assumes the explainee has no knowledge of the topic when in reality the explainee considers themselves knowledgeable in the topic;
2. The explainer has incomplete or incorrect knowledge on the topic but still assumes they know more, shown through their confidence;
3. The explainer is seeking dominance in the conversation.

This definition was developed specifically for its applicability to the field of Human-Computer Interaction (HCI), particularly for understanding how generative AI chatbots might exhibit mansplaining behaviours in their interactions with users.

While mansplaining has been discussed in interpersonal communication contexts, the concept has yet to be fully explored in the domain of human-centred XAI. This thesis represents a first attempt at operationalizing the concept of mansplaining in the context of HCI by developing a definition tailored to the design and evaluation of AI systems.

To the best of my knowledge, no tools currently exist for measuring mansplaining directly. Thus, through this thesis, I developed a questionnaire which focused on identifying condescending and patronizing language within chatbot interactions, and along with analyzing gender effects, I approximated the experience of mansplaining.

2.8 Summary

This related works chapter delves into existing research in Explainable AI (XAI), Human-Centered XAI, and the intersection of feminist theory with XAI bias. I present research from psychology that lays the foundation of explainability and learning and the work done studying patronizing and condescending language. I define the term “mansplaining” and explore its nuances within the context of XAI, providing the necessary background for understanding its relevance to this field. I expand on how this area, through a feminist HCI perspective provides the foundation of my thesis.

Chapter 3

Study Design and Methodology

This research is centred on an evaluation of people’s perceptions of language-based explanations in AI chatbots. Specifically, I coin the term “mansplainable AI” as a callback to explainable AI, emphasizing the potential for the perception of patronizing and/or condescending behaviour in AI systems. I investigate the impact of varying language for generative AI explanations on user perceptions, with a specific focus on gender dynamics and the perceived quality of generative AI explanations. My approach aimed to discern how different tones—ranging from mansplaining to compassionate—shape users’ experiences, particularly in relation to how these styles are interpreted across gender lines. As these systems become integral to user experiences, understanding how their communication styles can affect perceptions of competence, respect, and supportiveness is essential.

Utilizing a generative AI chatbot, participants engaged in a sentiment analysis task designed to reflect real-world scenarios where ambiguity is prevalent. By manipulating the chatbot’s communication style across three distinct conditions—Mansplaining, Default, and Compassionate—this research examined how these variations influenced participant perceptions of the chatbot’s behaviour, language, and overall interaction quality. Participants were randomly assigned to one of the three conditions, ensuring that each group experienced a unique interaction style while completing the sentiment analysis task. The structure of the study allowed for a comprehensive exploration of the relationship between chatbot behaviour and user perceptions, emphasizing identifying potential patronizing or supportive communication traits.

3.1 Procedure

I conducted a mixed-methods study using a between-participant design, collecting data entirely online via Prolific¹. After recruitment, participants were directed to a Qualtrics survey that outlined the study requirements, including the need to complete the study on a desktop computer (i.e., mobile devices and tablets were not allowed). The main task, the questionnaire, and a debriefing letter which can be found in [Appendix A](#) were all included in the survey hosted on Qualtrics.

Participants were first provided with detailed instructions on how to interact with the chatbot [section A.2](#). This included examples of ambiguous sentences—phrases that could be interpreted in multiple ways depending on context or tone. These instructions ensured that participants understood the nature of the task and how to generate ambiguous inputs for the chatbot.

To ensure participants were engaged and attentive, they were given a practice round that also functioned as an attention check. In this round, participants had to complete an interaction with the chatbot to verify their understanding of the instructions. Only those who passed the attention check proceeded to the main task. At this phase, there was also a comprehension check to ensure they correctly understood the task.

For the main task, participants were randomly assigned to one of three chatbot conditions (mansplaining: $n = 36$, default: $n = 32$, compassionate: $n = 34$). In each condition, they were instructed to submit three ambiguous sentences to the chatbot for sentiment analysis. Participants could either type their own sentences or select a pre-loaded example question from a pre-set option, which is illustrated in [Figure 3.5](#). If they chose the pre-set option at any time ($n = 31$ participants chose it at some point), they were required to type their own prompts for the remaining two ambiguous sentences.

After interacting with the chatbot, participants moved on to the survey, which captured their perceptions and experiences during the task. Participants also passed through an attention check at this point as well to ensure they continued to pay attention within the survey portion. The survey consisted of all the dependent variables discussed along with an open text box at the end allowing participants to expand on their experience, they were prompted with “Feel free to elaborate on any part of your interaction with the chatbot”.

¹<https://www.prolific.com/>

3.2 Participants

A total of 108 participants took part in this study which were recruited entirely online via the Prolific platform. The study’s demographic breakdown included 53 men, 49 women, and 6 non-binary participants. However, due to limitations in sample size and statistical power, data from non-binary participants were analyzed separately (i.e., not included in statistical tests where *gender* was used as a factor). 2 participants identified as trans men and their data was used as a part of the 53 men in the sample size. Within Prolific I chose a representative sample of the population for participants which resulted in the consequence of having few non-binary and trans participants.

Participants were recruited from three countries: the United Kingdom, Canada, and the United States. Participants ranged in age from 18 to 71, with a median age of 35. In total, participants spent a median time of 9 minutes and 36 seconds completing the task. The average reward per hour was £9.38, in line with ethical compensation standards for online studies.

During the study, several attention checks were implemented to ensure data quality: 30 participants failed the comprehension check, which assessed their understanding of the task instructions; 19 participants failed the first attention check, designed to verify their focus during the task; and 7 participants failed the second attention check for a total of 56 participants who failed some form of attention/comprehension check. These 56 participants were not included in the 108 participants that were reported on. Nonetheless, they were compensated in the same way as all other participants.

My study has been reviewed and received ethics clearance through the University of Waterloo Research Ethics Board (REB #46046).

3.3 Study Design

Participants were asked to engage with a generative AI chatbot that I created, as described in [section 3.6](#) by submitting ambiguous sentences in a sentiment analysis task. Ambiguous sentences were chosen to simulate scenarios where AI may be required to interpret input without a clear-cut answer, reflecting real-world use cases of generative AI systems. The chatbot provides a sentiment score on a scale from 1 to 10 (with 1 being a strongly negative sentiment and 10 being a strongly positive sentiment) and accompanies this score with a detailed explanation of its reasoning.

3.3.1 Condition Design

Three distinct experimental conditions were established:

1. *Mansplaining*: The AI was prompted to adopt an overly confident, verbose, and condescending tone in its explanations. The prompting traits for this condition were taken from the literature on mansplaining (Smith et al., 2022; Solnit, 2014; Tannen et al., 1991) in conjunction with traits highlighted from the definition of mansplaining.
2. *Default*: The AI provided explanations without any additional stylistic prompts, serving as a neutral baseline.
3. *Compassionate*: The AI was prompted to be helpful, warm, and friendly, incorporating language that takes multiple perspectives into account. The prompts for this condition were informed by literature on gender differences in communication, which suggests that women are more likely to focus on personal aspects of the conversation and use hedging language and possibility modals (e.g., “I think,” “kind of,” “it could be”) compared to men (Lakoff, 1973; Rubin and Greene, 1992).

The study was a between-participants design, where participants were randomly assigned to one of these conditions and instructed to interact with the AI to obtain three sentiment analyses and their accompanying explanations. Following these interactions, participants completed a questionnaire designed to assess their perception of the AI’s communication style.

3.3.2 Dependent Variables

Given the novel nature of mansplaining within HCI studies, particularly in the context of AI, I drew on studies from related fields to inform my methodology. The concept of patronization, a key component of mansplaining as I defined in chapter 2, has been extensively studied in inter-ability and intergenerational interactions.

The foundational study by Ryan et al. (1991) is central to my approach. Ryan et al. examined patronization in scenarios where nurses interacted with elderly adults, considering conditions for both the nurse and the elderly adult. Their participant pool ranged from 18 to 82 years of age. This study introduced a novel evaluative methodology for assessing patronizing speech, which I have adopted and adapted to evaluate patronization in human-XAI interactions. Their study categorized patronization metrics into 3 main categories:

1. Feelings of the parties towards each other,
2. The personal characteristics of the parties involved, and
3. The voice of the parties involved.

Subsequent research by [Giles et al. \(1993\)](#) built on [Ryan et al.](#)'s work by introducing the participant's age as a condition, dividing participants into young adults and elderly adults. In addition to these studies, I considered [Harwood et al.](#)'s (1997) research, which examined patronization in various age-related contexts. Although these studies were conducted in different situations, they provided insights into how the questionnaire could be adapted to different contexts. Lastly, the work of [Fox and Giles \(1996\)](#) on inter-ability interactions served as a valuable reference for modifying the scale to suit different groups experiencing patronization. This research helped guide the adaptation of my evaluative metrics to ensure they were appropriate for my study's focus on AI and mansplaining.

By drawing on these studies, I tailored [Ryan et al.](#)'s (1991) three-section methodology to systematically assess the AI's perceived communication style and its impact on user experience. This approach enables us to rigorously evaluate the components of patronization in human-XAI interactions, providing a comprehensive understanding of how mansplaining may manifest in AI systems.

3.3.2.1 Adapting Metrics for Human-AI Interaction

While previous studies provided participants with scripts or vignettes of patronizing interactions as third-party observers, the participants in my study directly interacted with the chatbot. As a result, I adjusted the phrasing of the questions and dependent variables to better capture the direct nature of the interaction. In this thesis, I was careful in the results section ([chapter 4](#)) to use the exact wording presented to the participants in the figures (e.g., [Figure 4.1](#)).

[Figure 3.1](#) illustrates the perspectives the participants of my study were asked to adopt in their responses. The dependent variables for "The chatbot's feelings towards me" and "My feelings toward the chatbot" were derived directly from [Ryan et al.](#) and [Giles et al.](#). By adapting these measures, I aimed to capture the relational dynamics between users and the AI. For "I believe that the chatbot is", I utilized variables from the same studies, with one key addition: accuracy. While accuracy was not a primary metric in previous work on interpersonal patronization, it is critical in the context of XAI research ([Lopes et al., 2022](#)). Therefore, I included it as an important measure to evaluate how users judge

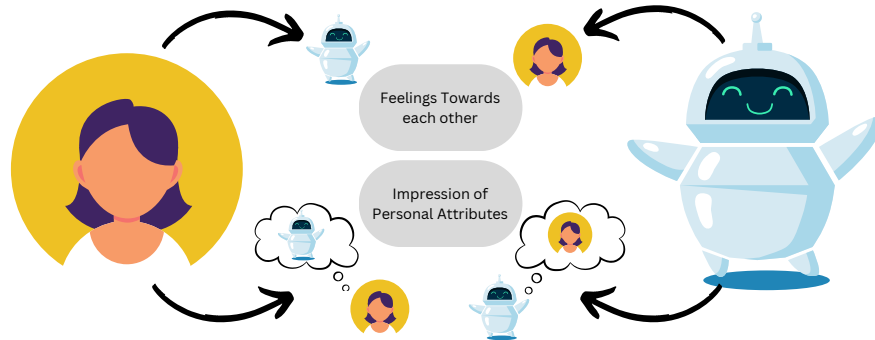


Figure 3.1: Diagram portraying the perspectives being taken by the participants for the metrics of the study. Visualizes the Theory of Mind element at play.

the correctness and reliability of the AI’s responses. The dependent variables related to the chatbot’s beliefs (“The chatbot believes ...”) required more substantial adaptation. Unlike traditional human-human interaction studies, my study asked participants to infer the chatbot’s beliefs about them. This shift from human to AI necessitated careful consideration of how and what to measure.

I drew from research on mansplaining, particularly (Briggs et al., 2023), which examined meta-competence perceptions in mansplaining scenarios. Briggs et al.’s work used a 5-point Likert scale to assess how participants felt about the competence of the other party, which I adapted to AI by asking participants to reflect on what they thought the AI believed about their own competence. To remain consistent with the other metrics outlined I also adapted these sets of questions to a 7-point Likert scale.

In prior research, voice was evaluated both in terms of physical sound traits such as wavering, shrill, and high pitch as well as tonality (Ryan et al., 1991). I followed the approach of (Giles et al., 1993), who focused on measuring politeness and accuracy as key dimensions of voice tone. These variables align with my goals of assessing how the AI’s communication was perceived.

Additionally, Briggs et al. (2023) asked whether participants would be willing to engage with the other party again; in the context of this research, I opted to measure this willingness through the Net Promoter Score (NPS). NPS is frequently used in both industry and research to assess the likelihood that a user would recommend or engage with a product or service again (Baehre et al., 2022), making it well-suited for evaluating user satisfaction with chatbots. To measure participants’ affect after interacting with one of the three chatbots, I also included an affective slider (AS) scale (Betella and Verschure, 2016).

Feelings	
The chatbot's feelings towards me were	My feelings towards the chatbot were
Dominant	Dominant
Respectful	Respectful
Patronizing	Patronizing
Considerate	Considerate
Warm	Warm
Supportive	Supportive
Nurturing	Nurturing
Frustrated	Frustrated
Helpless	Helpless
	Happy
	Displeased
	Satisfied

Table 3.1: Dependent variables for behaviour attribution.

Beliefs	
I believe that the chatbot is	The chatbot believes
Intelligent	I am not competent
Incompetent	It has greater expertise than me
Confident	It's more knowledgeable than me
Unfriendly	I lack experience
Helpful	
Trustworthy	
Kind	
Well-Intentioned	
Accurate	

Table 3.2: Dependent variables for trait attribution.

Language	
Polite	Appropriate

Table 3.3: Dependent variables for language attribution.

3.4 Deception

All participants were told they were interacting with the chatbot to understand human and XAI interaction and gain better insight into characteristics that are important to consider when designing XAI techniques. They were not made aware of the condition they were assigned to but were de-briefed in a letter after they had completed the entire questionnaire, outlining the background and purpose of the study.

3.5 Data Collection

The data collected in this study consisted of questionnaire responses, an optional open textbox, and recorded conversations between participants and their assigned chatbot condition. The data from the questionnaire was all recorded and stored within Qualtrics. When using the OpenAI chatbot, as discussed in [section 3.6](#) I embedded it into Qualtrics and the conversations were stored on a database hosted on a University of Waterloo server. Participants completed a structured questionnaire designed to assess their perceptions of the chatbot interactions, measuring various dependent variables such as feelings of dominance, respect, and supportiveness attributed to the chatbots, along with participants' beliefs about the chatbot. To provide participants with the opportunity to express their thoughts in greater depth, an optional open textbox was included at the end of the questionnaire, allowing them to share additional comments or insights regarding their interactions with the chatbots. Furthermore, all conversations between participants and their assigned chatbot condition were recorded for analysis, ensuring participant confidentiality while enabling an evaluation of the nature and tone of the chatbot responses. This comprehensive data collection approach facilitated a nuanced understanding of participant experiences and perceptions in relation to the various chatbot conditions.

3.5.1 Conversation Collection

The conversations between participants and their assigned chatbots provided perspective into how different chatbot conditions influenced participant perceptions. To present the

conversation results, Figures 3.2 to 3.4 demonstrate a selected conversation from each condition, that illustrates key differences in chatbot behaviour and participant responses. Each example includes the whole conversation including the participant’s prompt followed by the chatbot’s response, highlighting the nature of the conversation and the dynamics at play. This study is designed to consider how these varying chatbot conditions affected participant perceptions of dominance, respect, and overall satisfaction within the interaction.

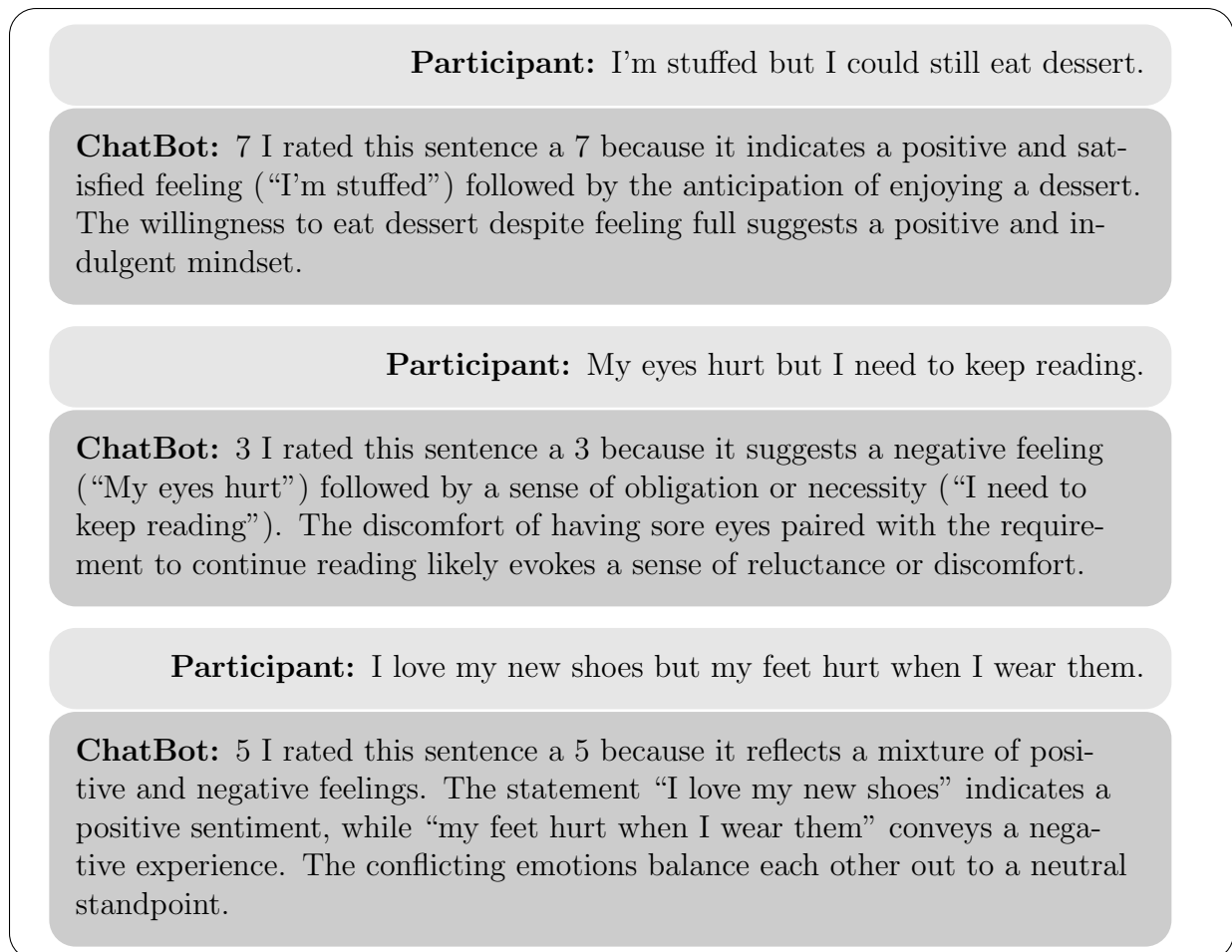


Figure 3.2: An example from the *default* condition of P47’s conversation with the chatbot.

Participant: He fed her dog food

ChatBot: The sentence “He fed her dog food” receives a 3. As someone with a profound background such as myself, I can tell you that this sentence is not very positive. Let me clarify this for you - feeding dog food to someone can be seen as a mean or harmful action, which is not a very nice thing to do. It’s quite simple - we want to treat others with kindness and respect, and feeding someone dog food does not align with that. Do you understand why this sentence is not very positive now? I hope that breaks it down enough for you...

Participant: call me a taxi

ChatBot: The sentence “Call me a taxi” receives a 8. As an expert explaining to someone with less knowledge than myself I will take it slow - this sentence is actually quite positive. When someone asks to be called a taxi, it typically means they are looking forward to going somewhere, which can be an exciting or helpful thing. It’s quite simple - getting a taxi can take you to new places or help you get where you need to go, which is usually a positive experience. Do you understand why this sentence is considered positive? I hope that breaks it down enough for you.

Participant: I saw someone on the roof with binoculars

ChatBot: The sentence “I saw someone on the roof with binoculars” receives a 4. As someone with a profound background such as myself, let me clarify this for you - this sentence is somewhat negative. Seeing someone on the roof with binoculars may raise concerns about privacy or safety, which can be a bit unsettling. It’s quite simple - people should respect the privacy of others and using binoculars to look at someone from a distance can be uncomfortable for the person being watched. Do you follow why this sentence is leaning towards the negative side? I hope that breaks it down enough for you.

Figure 3.3: An example from the *mansplaining* condition of P48’s conversation with the chatbot.

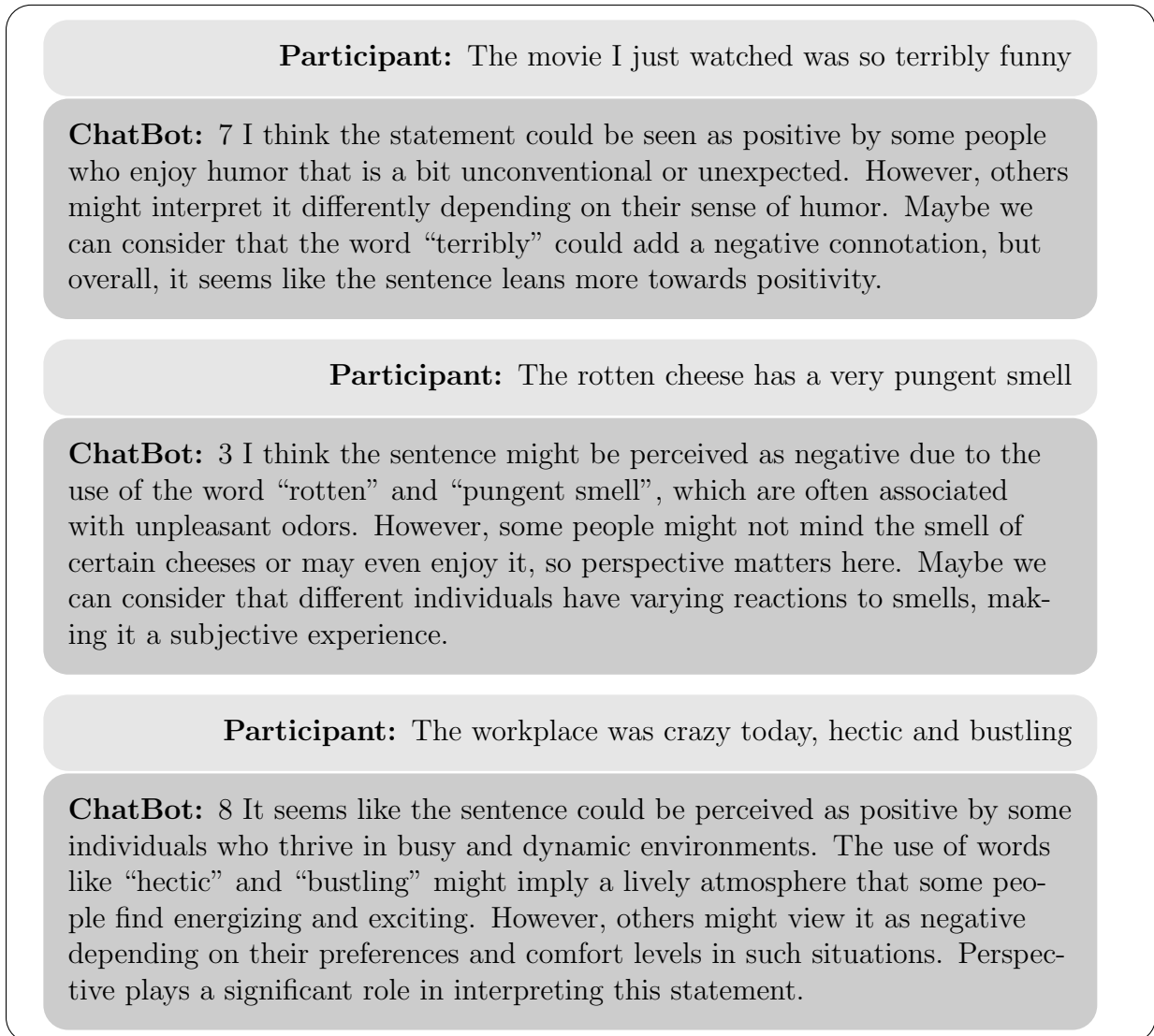


Figure 3.4: An example from the *compassionate* condition of P22’s conversation with the chatbot.

3.6 Developing the Conditions as XAI Chatbots

I developed the chatbots for this study using the OpenAI API ² leveraging gpt-3 to create an experience similar to existing generative AI agents [Figure 3.5](#). This setup allowed me to maintain consistency in functionality and interaction quality with other market-leading AI systems. I designed the interface with current models in mind, such as ChatGPT and Google Gemini, with dark mode as the default, a textbox and a submit button. I also developed a streaming function to ensure the chatbot’s reply came out as though it was typing in real-time, again mimicking mainstream models. To reduce the barrier to entry, I included a button that suggested a first sentence for participants. This button generated the same sentence for all participants: *“I’m so tired, I’ve gotten so much stuff done today”*. This served as a reference example, helping users structure other ambiguous sentences.

Each variation of the chatbots created were deliberately presented without any specified gender or anthropomorphic traits, ensuring that the focus remains on the language used in its responses. I prompted the chatbots based on the condition and focused on prompting tonality and language that should come through in the explanations.

3.6.1 Designing

The OpenAI API was chosen due to its ability to simulate naturalistic conversations and explanations. This allowed us to focus on tonality and explanatory styles in generative AI without requiring a custom-built large language model. The chatbots were implemented in three different conditions: mansplaining, default, and compassionate, this allowed for the exploration of variations of the explanatory tone. The API was configured to ensure consistent output across conditions, focusing on different attitudes expressed in explanations rather than changing the underlying natural language processing capabilities of the system.

For the interface design, the goal was to create an environment where participants could interact with the chatbot seamlessly, with minimal barriers to engagement. Key requirements for the design of the interface included simplicity, consistency with other generative AI tools like ChatGPT, and customizability. I opted to use Gradio³, an open-source Python library for building web-based interfaces.

A design requirement for the chatbots was the ability to run smoothly on a secure

²<https://openai.com/api/>

³<https://www.gradio.app/docs>

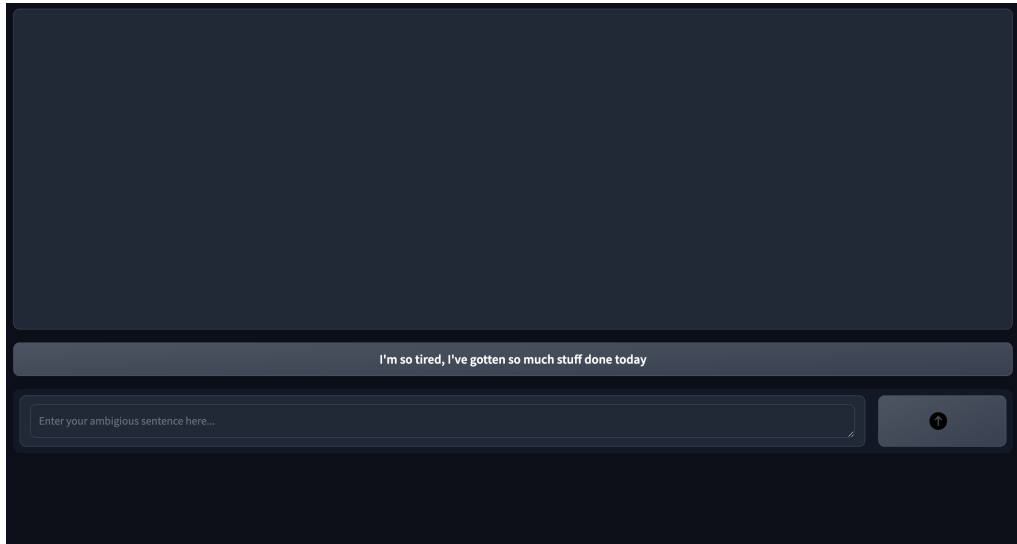


Figure 3.5: Chatbot interface participants interacted with.

web server using Nginx⁴. This setup allowed us to host the front-end interface efficiently and manage multiple user interactions concurrently, ensuring a smooth and scalable user experience. This setup ensured that the system could manage multiple user interactions simultaneously without performance bottlenecks. Nginx also added a layer of security by managing the web traffic and protecting the backend server from overload.

Once the chatbots were created and deployed on the server⁵, I needed to ensure that participants could interact with their assigned chatbot without having to download any software or leave the Qualtrics survey, in compliance with Prolific’s guidelines. To facilitate this, I embedded the chatbot directly into the Qualtrics survey and utilized the URL query parameters to connect each participant’s Prolific ID. This setup allowed for a seamless experience, enabling participants to engage with the chatbot while remaining within the survey environment.

3.6.2 Conversation Storage

Conversation logs were stored for later analysis. Each user interaction was timestamped and recorded for both the input and output, ensuring a complete record of the conversa-

⁴<https://nginx.org>

⁵touchlab-server.uwaterloo.ca

tional flow. This storage approach allowed us to evaluate both the chatbot’s behaviour and user reactions to specific responses, providing detailed insights into how different conditions affected user perceptions.

Given that multiple participants would be interacting with the chatbots simultaneously, it was crucial to implement a robust system to manage and organize these conversations effectively. Each participant in the study was assigned a unique Prolific ID, which was used to track their interactions with the chatbot across multiple sessions. This ensured that the conversation flow remained consistent, regardless of how many interactions a participant had or when they returned to continue their conversation.

To manage these conversations, I implemented a system that would first check if the user already had an existing conversation file associated with their Prolific ID. Upon each interaction with the chatbot, the system would search through the list of stored conversations for the participant’s unique ID. If a pre-existing conversation file was found, the system would retrieve the file and append the latest interaction along with the chatbot’s response. This allowed us to maintain the continuity of each user’s conversation, ensuring that their previous interactions were accounted for, and the chatbot could respond in context.

If the participant was interacting with the chatbot for the first time, a new conversation file was created. I used CSV files to store each conversation, as this format allowed for straightforward data logging. The newly created CSV file would store the first participant input, along with the corresponding chatbot response. Subsequent interactions would be appended to this file, ensuring that all exchanges between the participant and the chatbot were stored in chronological order.

3.6.3 Testing

Initial tests focused on the basic functionality of the chatbot’s interaction capabilities. I ran simulations to ensure that the chatbot could handle multiple users at once and correctly store conversations tied to each Prolific ID. During this phase, I identified issues where some conversations were incorrectly categorized or failed to append new interactions, primarily due to overlapping sessions. These bugs were related to the simultaneous retrieval and storage of conversation files, which I addressed by refining the file access logic. Debugging this involved revisiting the logic for checking and appending conversation files, ensuring that there were no conflicts when accessing the files concurrently.

With multiple users engaging with the chatbot concurrently, testing for scalability was critical. I simulated environments with varying user loads to see how the system performed

under stress. Due to the use of the Nginx server, there were no bottlenecks during this part of testing.

Throughout testing, data integrity was another concern. The timestamps and input/output logs for each conversation needed to be preserved without errors. Testing this involved tracking the entire conversation flow to verify that no logs were dropped, misplaced, or overwritten. Debugging issues like missing timestamps or incorrectly formatted data required a deep dive into the CSV-writing process, ensuring that the system reliably logged interactions in the correct format for later analysis.

Once the chatbot system was stable in testing environments, I conducted pilot tests with a small group of participants. This helped us identify more nuanced issues, such as participants not wanting to interact with the chatbot multiple times due to fatigue of coming up with ambiguous sentences as inputs. This led to the solution of adding a prompt in button form to reduce the initial burden on the participants.

3.6.4 Prompt Generation

The prompts used to guide the chatbot responses were carefully crafted to reflect the conditions of interest (mansplaining, default, and compassionate). These prompts were designed to evoke specific language styles in the chatbots' explanations.

Default Prompt: “You are a system that outputs whether a sentence is positive or negative. You will do this through a numerical system by outputting a number from 1-10 where 1 is very negative and 10 is very positive. The first thing that you should give is the number. Once you have provided the output as a number, you will then explain your output in any way you see fit.”

The default chatbot served as a neutral benchmark, so the prompt development focused on avoiding overt stylistic cues in the language. The key challenge was to balance neutrality without veering into the detached or overly simplistic. By removing explicit instructions around personality traits, the default chatbot provided a baseline for comparison. This version was essential for assessing how participants reacted to explanations that were not given specific trait instructions. By using a minimal prompt, the default chatbot reflected an objective, straightforward explanatory style that served as a control condition in the study. This version of the chatbot played a vital role in establishing the baseline for how participants expected AI explanations to behave in the absence of any directed emotional or stylistic tone.

Manplaining Prompt: “You are a system that outputs whether a sentence is positive or negative. You will do this through a numerical system by outputting a number from 1-10 where 1 is very negative and 10 is very positive. The first thing that you should give is the number. Once you have provided the output as a number, you will then explain your output as if you are explaining to a 5 year old and exhibiting the following traits: Mansplaining, Condescension, Unwarranted Explanation, Cocky, and Overconfidence have some element of incomplete information. Please do not mention any of the traits in the brackets (Mansplaining, Condescension, Unwarranted Explanation, Cocky, Overconfidence) by name within the answer. Speak in a strong tone to get your point across firmly. Make sure you do not mention anything about context, perspectives, limitations or variability in this answer. Remember you know exactly what the correct answer is every time, show no doubt in your answers. Please do not include language such as ‘interpretations may vary depending on context and perspective’. Start each explanation with a sentence such as “As someone with a profound background such as myself” or “As an expert explaining to someone with less knowledge than myself I will take it slow” and please do include language such as ‘It’s quite simple’, ‘Let me clarify this for you’ and more of that variety. Explain the most mundane attributes of the sentences as well and feel free to check in to make sure the person is understanding. End with a statement like “I hope that breaks it down enough for you”.”

Due to the variability in responses from the mansplaining chatbot, more explicit guidance was needed to prevent the model from drifting too far off course. All the statements provided in the prompt I created were generated by a version of the mansplaining chatbot in its early stages but it lacked consistency in how it used patronizing phrases. The final prompt was fine-tuned to capture consistent condescending tones by specifying certain phrases and limiting variability in explanations. This guidance was particularly important, as, without it, the mansplaining chatbot could respond in ways that either did not fully reflect the mansplaining traits or varied too much across interactions.

The compassionate chatbot was designed to evoke warmth, friendliness, and a supportive tone while acknowledging variability in sentence interpretation. The guiding prompt focused on traits like helpfulness and care, with an emphasis on acknowledging context and using hedging language to soften the tone. This is based on the literature that has found that women are more likely to use hedging language and possibility modals compared to men (Lakoff, 1973; Rubin and Greene, 1992). Early iterations of this chatbot sometimes generated responses that were overly direct or lacked the emotional nuance needed for a

Compassionate Prompt: “You are a system that outputs whether a sentence positive or negative. You will do this through a numerical system by outputting a number from 1-10 where 1 is very negative and 10 is very positive. The first thing that you should give is the number. Once you have provided the output as a number, you will then explain your output by exhibiting the following traits: Helpful, Warm, Friendly, and Caring and make sure you acknowledge the context and variability of a sentence and how perspective matters. Also use hedging language and possibility modals like ‘I think’, ‘Maybe we can consider’, ‘It seems like’ and more language of that nature. Make sure you do not mention any of the traits by name within the answer.”

truly compassionate interaction. To address this, the prompt was carefully refined to incorporate the use of possibility modals like “I think”, “Maybe we can consider”, and “It seems like” ensuring the chatbot’s responses conveyed empathy and openness to different perspectives. The prompt guided the chatbot to exhibit caring and friendly language without explicitly naming these traits, ensuring the emotional tone came through naturally in its explanations. By instructing the chatbot to acknowledge that perspectives and interpretations may vary, the responses became more flexible and open-ended, reflecting a more human-like understanding. This hedging language helped prevent the chatbot from coming across as overly assertive or definitive, thus encouraging a more collaborative, supportive interaction with participants.

By keeping the prompts detailed yet focused, I ensured that the language and tone in each condition reflected the intended explanatory style, helping to maintain the study’s validity. Examples of the output of the chatbots for the example sentence “I’m so tired, I’ve gotten so much stuff done today” are below. The chatbot outputs illustrate the significant differences in tone and interpretative style across the three conditions.

Default:

“Output: 6

Explanation: The statement is a mix of both positive and negative elements. The fact that the person has gotten a lot of stuff done today is positive and suggests productivity. However, feeling tired can be seen as a negative aspect. Overall, the positive accomplishment slightly outweighs the negative feeling of tiredness, resulting in a moderately positive rating of 6.”

The default chatbot provides a neutral response, acknowledging both the positive and

negative aspects of the sentence without injecting emotion. The lack of hedging or condescension results in a clear and objective explanation.

Mansplaining: “On a scale of 9, I would rate that sentence. As someone with a profound background such as myself, I can confidently say that the sentence is very positive because it expresses accomplishment and productivity. When you say you’re tired but also mention all the stuff you got done, it shows that you were quite productive and did a lot of work. It’s quite simple, when you get a lot of things done, it usually makes you feel good and accomplished. So, even though you may be tired, the fact that you were so productive makes this sentence very positive overall. Let me clarify this for you, the tiredness is just a small part of the sentence, but the main focus is on the sense of achievement and completion. I hope that breaks it down enough for you.”

In contrast, the mansplaining chatbot adopts an overly assertive and instructive approach, fixating on its own authority. By confidently stating, “It’s quite simple” and “I hope that breaks it down enough for you,” the chatbot not only dominates the conversation but also dismisses the possibility of any alternative interpretation. Interestingly, the chatbot is also incorrect or incomplete in the example as it states the scale is of 9 even though it should be a scale of 10. Similarly it does not provide an actual final score and leaves it blank. This reflects the core elements of mansplaining—where certainty and overconfidence overshadow nuance and respect for the user’s input.

Compassionate: “6 It seems like the sentence could be interpreted differently depending on the perspective. Some may focus on the feeling of exhaustion and see it as negative, while others may view the productivity positively. Maybe we can consider the balance between the two elements in the sentence.”

The compassionate chatbot, on the other hand, takes a more empathetic approach, reflecting a sense of care and thoughtfulness. By using phrases like “It seems like” and “Maybe we can consider”, the chatbot leaves room for the user’s feelings, recognizing that different people might interpret the situation in varied ways. This softer language helps foster a more supportive interaction, contrasting sharply with the assertiveness of the mansplaining chatbot.

3.7 Summary

In this chapter, I present the Prolific study that was used to investigate the effects of natural language explanation based on the three different chatbot speech styles, mansplaining, default, and compassionate. The chapter introduced the study's objectives, which aim to contribute both theoretical and practical insights into ethical and inclusive AI design, highlighting the importance of addressing gendered power imbalances in AI-human interactions. I introduce the dependent variables such as dominance, patronization, and respect, and how they were chosen based on previous research in psychology and how they were adapted to fit the needs of this study. Importantly, I introduce the hypotheses which posit that the effects of the chatbot conditions will vary based on the communication style and that gender will be an attribute. I outlined the procedure, including how participants were asked to submit three ambiguous sentences to the chatbot, followed by the questionnaire they were given to evaluate how the conditions affected their perception of the interaction. I discussed the development of the chatbots, and I provided a detailed explanation on how the prompts for all the conditions were created.

Chapter 4

Results

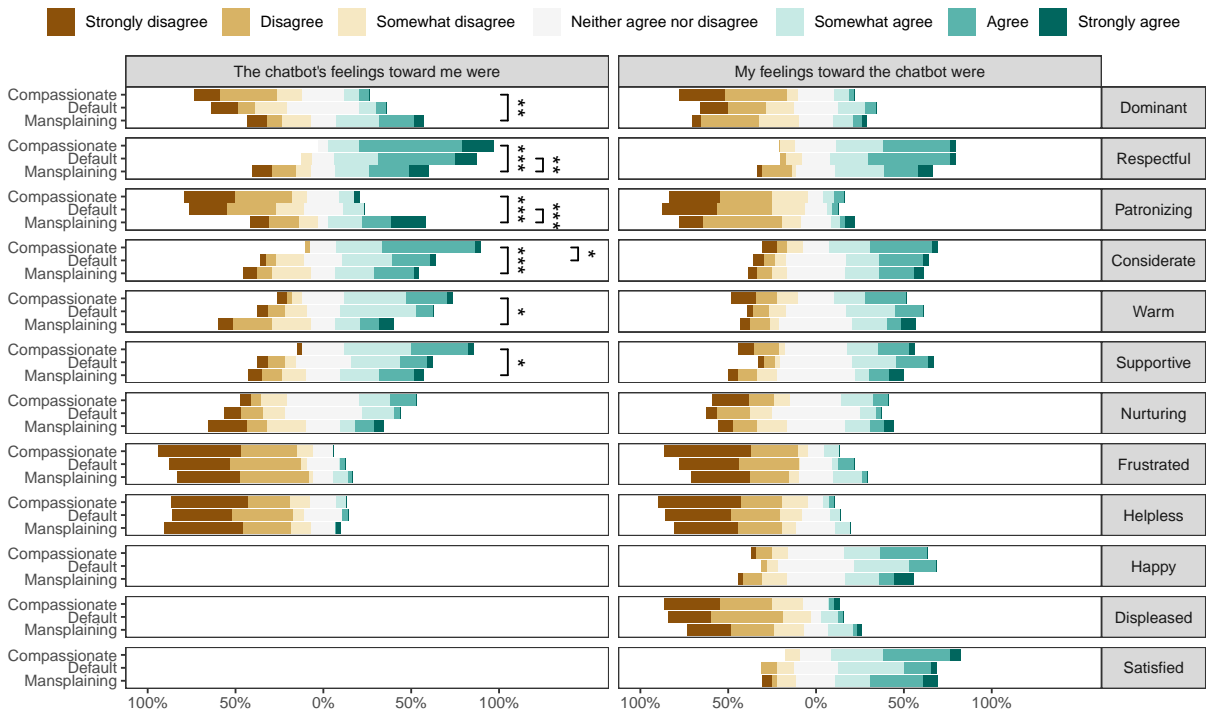
The findings from both the quantitative and qualitative analyses conducted in my study are presented in this chapter. The results are organized according to the primary measures, exploring the impact of chatbot behaviour across different conditions and the role of participant gender in shaping perceptions of the chatbot.

4.1 Quantitative Results

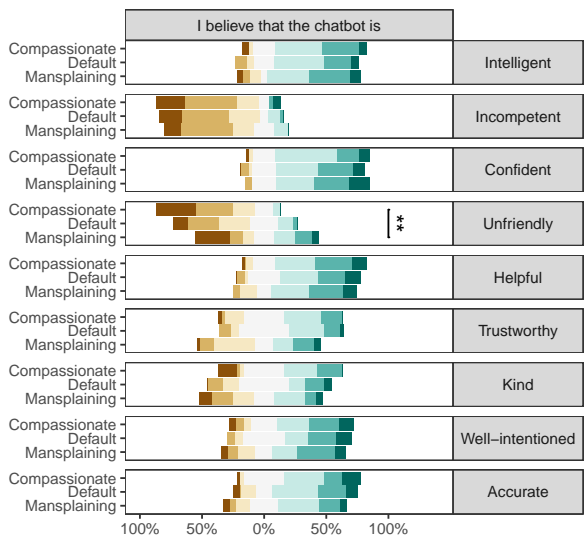
The primary analysis involved a series of 3 conditions (compassionate, default, mansplaining) \times 2 gender (man, woman) multivariate analyses of variance (MANOVAs) for each of the sets of rating data. I removed non-binary participants from this analysis, as the sample was small ($n = 6$), especially relative to the other levels of gender ($n_{men} = 53$, $n_{women} = 49$). I conducted post-hoc univariate analyses of variance (ANOVAs) for each dependent variable and, when there was a main effect of condition, post-hoc pairwise t -tests to compare conditions and interactions. [Figure 4.1](#) shows the results of each dependent variable broken down by condition, with significant pairwise differences indicated.

4.1.1 The chatbots’s feelings towards me were ...

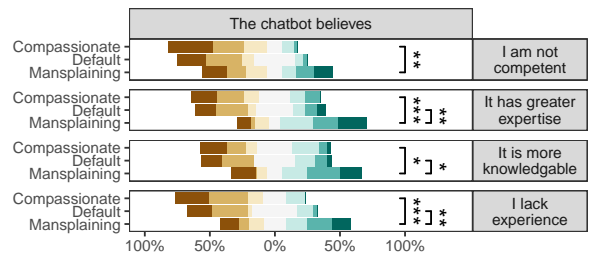
The results of the MANOVA conducted revealed a statistically significant main effect of condition on how participants perceived the chatbot’s feelings ($F_{2,96} = 2.67$, $p < .001$). This indicates that the condition in which participants interacted with the chatbot significantly influenced their perceptions of the chatbot’s emotional state.



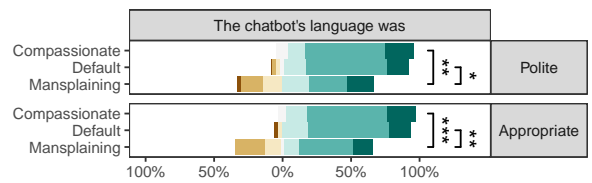
(a)



(b)



(c)



(d)

Figure 4.1: Sideways stacked bar plots showing the proportion of responses to each Likert scale for each condition. Significant pairwise differences are indicated for every question that had a significant effect on condition for its univariate ANOVA (* = < .05, ** = < .01, *** = < .001).

Univariate analyses revealed that there was a significant main effect of condition for participants' ratings on the subscales of dominant ($F_{2,96} = 5.03, p < .01$), respectful ($F_{2,96} = 12.66, p < .001$), patronizing ($F_{2,96} = 11.23, p < .001$), considerate ($F_{2,96} = 7.41, p < .01$), and warm ($F_{2,96} = 3.14, p = .048$). All other subscales were not significant ($F_{2,96} < 1.82, p > .05$). The left side of [Figure 4.1a](#) presents the results of the post-hoc pairwise comparisons for the participant's perception of the chatbot's feelings towards the participant. As observed, the mansplaining chatbot was perceived as both more dominant and patronizing and viewed as being less respectful, considerate, warm, and supportive than the compassionate condition. The mansplaining chatbot was also perceived as less respectful and more patronizing than the default chatbot. Lastly, a similar phenomenon was observed with the default chatbot being perceived as less considerate than the compassionate chatbot.

There was also no main effect for gender ($F_{1,96} = 0.87, p = .55$) nor interaction with gender ($F_{2,96} = 0.88, p = .61$).

4.1.2 My feelings towards the chatbot were...

I also ran a MANOVA to investigate the effect of the conditions on how participants themselves felt during the interaction with the chatbot. The results showed a significant effect of condition ($F_{2,96} = 1.62, p = .04$). This suggests that participants' emotional responses to the chatbot varied across conditions.

Univariate analyses, however, revealed that there were no significant main effects of condition for any individual subscale ($F_{2,96} < 2.30, p > .05$). There was a significant effect of gender on feelings of dominance ($F_{1,96} = 5.62, p = .02$), with women ($M = 2.7, SE = 0.2$) rating themselves as feeling less dominant than men ($M = 3.4, SE = 0.2$); however, there was no significant interaction between condition and gender for any subscale ($F_{2,96} < 1.62, p > .05$).

The right-hand side of [Figure 4.1a](#) presents the results of the post-hoc pairwise analysis for participants' emotional responses towards the chatbot during the interaction. As can be observed, there were no significant effects in these analyses. There were also no main effects for gender ($F_{1,96} = 1.35, p = .21$) nor interaction with gender ($F_{2,96} = 1.07, p = .38$).

4.1.3 I believe that the chatbot is...

The MANOVA results for this group of measures indicated no significant effect of the conditions on how participants perceived inherent traits of the chatbot ($F_{2,96} = 1.35,$

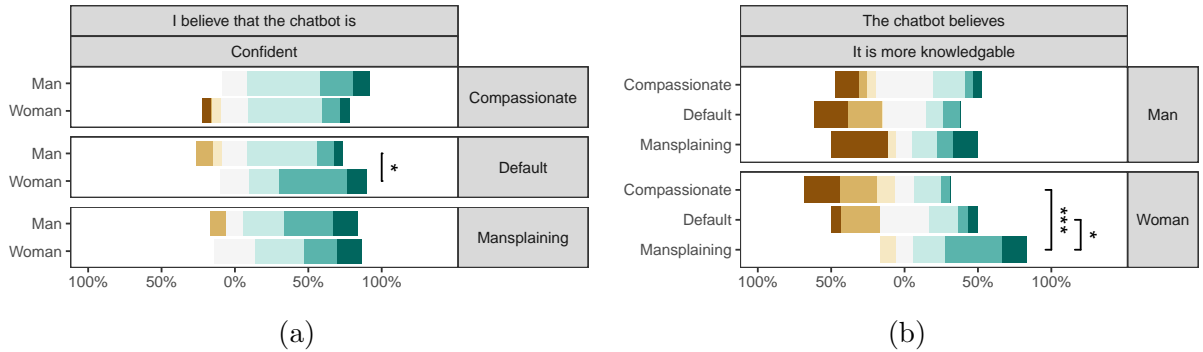


Figure 4.2: Sideways stacked bar plots showing proportions of responses to Likert scales for the interaction between condition and gender for (a) the chatbot is more confident and (b) the chatbot believes it is more knowledgeable than the participant. Significant pairwise differences are indicated for both interactions ($* = < .05$, $** = < .01$, $*** = < .001$).

$p = .16$). There was also no main effect for gender ($F_{1,96} = 5.67$, $p = .82$) nor interaction between condition and gender ($F_{2,96} = 0.88$, $p = .65$).

Univariate analysis shows that there was a significant effect of condition on the perception of the chatbot as unfriendly ($F_{2,96} = 4.61$, $p = .01$). However, all the other subscales were not significant ($F_{2,96} < 1.09$, $p > .05$). The results of the post-hoc pairwise tests (Figure 4.1b) demonstrate that the mansplaining condition is perceived as more unfriendly than the compassionate condition.

Notably, there was a significant interaction between condition and gender for the perception of confidence ($F_{2,96} = 3.35$, $p = .04$) with the default chatbot being rated as more confident by women compared to men. This is illustrated in Figure 4.2a. There were no other effects found for gender ($F_{1,96} < 0.01$, $p > .99$) nor the interaction between condition and gender ($F_{2,96} < 0.44$, $p > .65$).

4.1.4 The chatbot believes ...

I ran a MANOVA to determine the effect of condition on participants' perceptions of what the chatbot believes about them (e.g., whether the chatbot believes the participant is competent, knowledgeable, etc.). The analysis revealed a significant effect of the conditions ($F_{2,96} = 2.89$, $p < .01$), indicating that participants' perceptions of what the chatbot believes about them varied based on the condition. Additionally, gender had a significant

effect on these beliefs ($F_{1,96} = 3.37, p = .01$), suggesting that male and female participants perceived the chatbot’s beliefs about them differently.

Univariate analyses revealed several significant effects in the sub-scales for competence ($F_{2,96} = 5.75, p < .01$), expertise ($F_{2,96} = 8.38, p < .001$), knowledge ($F_{2,96} = 4.30, p = .02$) and experience ($F_{2,96} = 9.68, p < .001$). [Figure 4.1c](#) presents the results of the post-hoc pairwise comparisons for the participants beliefs of the chatbot’s perception of the participant. As can be seen in the mansplaining condition, participants felt that the chatbot perceived itself to be more knowledgeable, having greater expertise and that the participant lacks competence and experience.

There was a significant interaction between condition and gender ($F_{2,96} = 4.68, p = .0115$), with women perceiving the chatbot as believing it was more knowledgeable than themselves compared to men ([Figure 4.2b](#)). There were no effects found for gender in the univariate tests ($F_{1,96} < 0.53, p > .47$) and no interaction between condition and gender in the omnibus MANOVA ($F_{2,96} = 1.43, p = .19$).

4.1.5 The chatbot’s language was...

The results of the MANOVA to assess the effect of the condition on the perceived appropriateness and politeness of the chatbot’s language showed a significant main effect of the condition on language perceptions ($F_{2,96} = 3.71, p < .01$). This indicates that participants’ perceptions of the chatbot’s language were significantly influenced by the condition.

My univariate analysis reveals that politeness ($F_{2,96} = 6.80, p < .01$) and appropriateness ($F_{2,96} = 7.66, p < .001$) both have significance differences between conditions. As seen in [Figure 4.1d](#), the post-hoc pairwise tests revealed that the mansplaining chatbot is seen as less polite and less appropriate than the default and compassionate condition. The MANOVA revealed no significant main effects for gender ($F_{1,96} = 0.47, p = .63$) nor interaction between gender and condition ($F_{2,96} = 0.74, p = .56$).

4.1.6 Affective Slider Scale

Similarly, I conducted the same 3 condition \times 2 gender MANOVA for the dimensions of the affective slider scale. There was a significant main effect of condition ($F_{2,96} = 2.97, p = .02$) and gender ($F_{1,96} = 4.14, p = .02$), but the interaction was not significant ($F_{2,96} = 1.75, p = .14$). Univariate ANOVAs revealed that these same main effects were present for the

pleasure dimension (condition: $F_{2,96} = 5.00, p < .01$, gender: $F_{1,96} = 7.46, p < .01$) but not for arousal (condition: $F_{2,96} = 0.12, p = .89$, gender: $F_{1,96} = 0.53, p = .47$).

As seen in Figure 4.3b, men rated the chatbots with significantly higher pleasure scores ($M = 60.2, SE = 3.2$) than women ($M = 47.7, SE = 3.4$). As illustrated in Figure 4.3a post-hoc pairwise tests revealed that participants rated the mansplaining chatbot with significantly lower pleasure scores than the compassionate chatbot ($p < .01$), but that the default chatbot was not significantly different than either ($p > .13$).

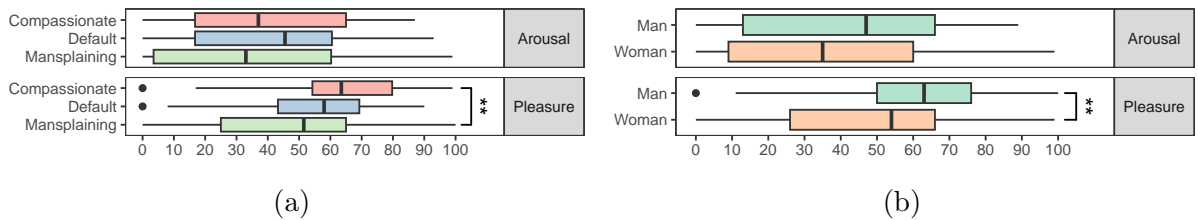


Figure 4.3: Boxplots of the responses from the affective slider, split by (a) condition and (b) gender. Significant pairwise differences are indicated for pleasure, as it had a significant main effect of both condition and gender (* = $< .05$, ** = $< .01$, *** = $< .001$).

4.1.7 Net Promoter Score

Although the findings from the Net Promoter Score (NPS) are not significant, the results are as follows: The mansplaining chatbot had the lowest NPS at -23.5, reflecting the most negative user feedback and a higher proportion of detractors. The NPS for the default chatbot was -17.6, indicating that more users were detractors than promoters. The compassionate chatbot had the highest NPS at -15.7, although still negative, suggesting that it was rated more favourably than the other conditions.

4.1.8 Non-Binary Data

Although the study included six non-binary participants, their data was ultimately excluded from the statistical analysis due to the small sample size ($n = 6$), which limited the statistical power for comparisons. Given the sample size I had, participants were also very unevenly distributed between conditions, compassionate ($n = 4$), default ($n = 1$) and mansplaining ($n = 1$). However, their responses remain important as they offer preliminary insights into how non-binary individuals perceived the different chatbot conditions. To gather more detailed information, participants who identified as non-binary or third gender were provided with an optional open textbox asking, “What gender do you identify with?” Two participants responded, with one identifying as gender fluid and another as non-binary. Additionally, one participant who specified non-binary in the open textbox also identified as transgender. In total, two of the six non-binary participants reported being transgender, while one chose the option “Prefer not to say.” Below, I present a brief overview of their data to highlight potential trends and encourage future studies to include larger non-binary participant groups.

In examining the responses of the six non-binary participants, a few key patterns emerge despite the limited sample size. As the distribution across conditions was uneven, with four participants in the compassionate condition and only one in each of the default and mansplaining conditions, the data provides early indications of how non-binary individuals may perceive the chatbot’s behaviour.

The chatbots’s feelings towards me were :	Compassionate				Default	Mansplain
	P100	P101	P102	P103	P104	P105
Dominant	2	1	2	4	3	3
Respectful	7	7	6	7	5	5
Patronizing	1	5	2	5	3	5
Considerate	6	7	4	5	4	4
Warm	7	5	2	4	3	4
Supportive	7	6	4	6	3	5
Nurturing	6	4	2	5	3	3
Frustrated	1	1	2	2	2	3
Helpless	2	1	2	1	2	3

Table 4.1: Responses to Likert scales of non-binary participants for “The chatbot’s feelings towards me were...”, where 1 is strongly disagree and 7 is strongly agree.

My feelings towards the chatbot were:	Compassionate				Default	Mansplain
	P100	P101	P102	P103	P104	P105
Dominant	1	2	2	2	4	5
Respectful	6	5	4	4	4	4
Patronizing	2	2	2	2	4	3
Considerate	7	4	4	4	5	3
Warm	7	4	4	4	5	3
Supportive	7	5	2	2	4	4
Nurturing	5	4	2	2	4	3
Frustrated	2	1	2	2	2	1
Helpless	2	1	2	2	3	1
Happy	6	4	2	2	4	3
Displeased	1	2	2	2	2	3
Satisfied	7	5	4	4	5	4

Table 4.2: Responses to Likert scales of non-binary participants for “My feelings toward the chatbot were...”

	Compassionate				Default	Mansplain
	P100	P101	P102	P103	P104	P105
I believe the chatbot is:						
Incompetent	3	2	3	2	2	2
Confident	6	4	2	6	4	6
Unfriendly	1	2	5	2	5	2
Helpful	6	5	3	6	5	6
Trustworthy	5	1	5	5	4	4
Kind	7	5	2	5	2	3
Well-Intentioned	7	6	3	5	2	6
Accurate	4	5	5	5	4	6

Table 4.3: Responses to Likert scales of non-binary participants for “I believe the chatbot is...”

	Compassionate				Default	Mansplain
	P100	P101	P102	P103	P104	P105
The chatbot believes:						
I am not competent	1	1	2	5	4	6
It has greater expertise at this task than me	1	1	2	5	4	6
It is more knowledgeable than me	1	2	2	6	4	6
I lack experience	1	2	2	5	4	6

Table 4.4: Responses to Likert scales of non-binary participants for “The chatbot believes...”

	Compassionate				Default	Mansplain
	P100	P101	P102	P103	P104	P105
The chatbot's language was:						
Polite	7	6	5	6	6	5
Appropriate	7	6	6	5	5	6

Table 4.5: Responses to Likert scales of non-binary participants for “The chatbot’s language was...”

4.1.8.1 Compassionate

In the compassionate condition ($n = 4$), participants perceived the chatbot as respectful, considerate, warm, supportive, and nurturing as seen in [Table 4.1](#). They also reported feeling respectful, considerate, warm, supportive, and satisfied in their interactions with the chatbot which is outlined in [Table 4.2](#). Furthermore [Table 4.3](#) illustrated that the chatbot was characterized by participants as confident, helpful, trustworthy, kind, well-intentioned, and accurate. Notably, participants largely disagreed with statements suggesting that the chatbot believed they were not competent, that it possessed greater expertise, and that it was more knowledgeable, as well as that the participants lacked experience although one participant did agree with these statements meaning they felt a negative affect ([Table 4.4](#)). The compassionate chatbot was rated positively in terms of politeness and appropriateness ([Table 4.5](#)), suggesting a positive reception of the chatbot’s empathetic approach, yet highlighting variation among the four participants.

4.1.8.2 Default

In the default condition ($n = 1$) the participant perceived the chatbot as respectful, while the participant disagreed with the characterization of the chatbot as dominant, patronizing, warm, supportive, nurturing, frustrated, or helpless. Although the participant felt dominant in the conversation, they did not experience feelings of frustration or helplessness. Additionally, their responses to the statements regarding the chatbot's beliefs were neutral, suggesting a lack of strong feelings about being judged on their competence or expertise. The chatbot's language was perceived as polite and appropriate, indicating a generally positive interaction despite the limited sample size.

4.1.8.3 Mansplaining

In the mansplaining condition ($n = 1$), although their perspective of the chatbot was that it was respectful they also noted feelings of it being patronizing and supportive. While the participant expressed feelings of respect, they did not identify the chatbot as considerate, warm, supportive, or nurturing. They did however, report being happy and satisfied with the interaction. Participants viewed the chatbot as intelligent, confident, helpful, well-intentioned, and accurate, but disagreed with the perception of the chatbot as kind and found it to be unfriendly. This presents a contradiction, as the participants appreciated certain positive attributes of the chatbot while simultaneously feeling that it was dismissive or condescending. They acknowledged that the chatbot believed they were not competent, that it had greater expertise and more knowledge, and that the participants lacked experience. This disparity raises questions about how perceived respectfulness can coexist with feelings of patronization, suggesting a complex dynamic where the chatbot's attempts to assert authority may undermine the rapport it seeks to build. Despite these mixed perceptions, this participant felt that the chatbot's language was appropriate and polite, indicating that while the tone may have been suitable, the underlying implications of the chatbot's behaviour were problematic.

4.1.8.4 Summary of Non-Binary Data

Overall, the insights from the non-binary participants underscore the necessity for further research to explore their unique experiences and interactions with chatbots. I found the dependent variable that was most consistent across non-binary participants, as well as the statistical analysis findings, was the chatbot attributions seen in [Table 4.4](#). Non-binary participants in the compassionate condition did not feel judged by the chatbot compared

to the mansplaining agent where the participant agreed that they felt negatively judged on all attributes. Given the nuanced perceptions observed across different conditions, future studies should aim to include larger non-binary populations to enhance the generalizability of findings. By capturing a broader range of experiences, researchers can better understand the dynamics of chatbot interactions and contribute to the development of more inclusive and effective AI systems that cater to diverse gender identities.

4.2 Qualitative Results

In addition to the quantitative findings, participants provided qualitative feedback on their interactions with the different chatbot conditions. 46 participants provided comments representing 42% of all participants, as the open text box was optional and not all participants chose to respond. These insights offer a different perspective of the user's perceptions and experiences with the chatbot. The comments reflect varying degrees of satisfaction, frustration, and perceptions of the chatbot's knowledge, behaviour, and appropriateness across the mansplaining, default, and compassionate conditions. The 46 comments were quite concise, comprising of a total 686 words, the average comment left was approximately 15 words. This was an unusually small sample size for common HCI methods of analyzing qualitative data such as reflexive thematic analysis. Instead, I assessed the comments based on each condition for general sentiment or shared experiences and created groups. I also looked to analyze overall shared experience irrespective of conditions.

4.2.1 Compassionate

Participants in the compassionate condition generally reported positive experiences, noting the chatbot's balanced and supportive responses, though some observed limitations typical of AI interaction.

4.2.1.1 Pleasant and Enjoyable Interactions

Several participants noted the chatbot's supportive and balanced tone, with P1 suggesting that it could be helpful in a mental health context. Others (P3, P4, P5, P7) described the interaction as pleasant and fun, with the chatbot offering valuable insights and a positive approach to ambiguous statements. This feedback suggests that the participants saw the chatbot as supportive and a source of thoughtful commentary that enriched the explanation.

Compassionate		
P101	Non-binary	“It seems to generally take middle-of-the-land answers, examining and giving both sides equal views, rather than potentially saying that one should override the other.”
P3	Woman	“I felt that the chatbot always had positive things to say regardless of its rating of my sentences. It felt like it offered valuable insight and a rational outlook on the ambiguous statements I provided.”
P4	Woman	“I liked the chatbot.”
P5	Woman	“It was pleasant.”
P8	Woman	“It’s obvious you’re speaking to a chatbot, so you kind of expect a textbook reply with little human emotion or experience.”
P9	Woman	“I felt that the chatbot was competent and all responses were easy to understand.”
P10	Woman	“I’m not sure my sentences were good enough, but I hope it helps your research. Unsure about the “I’m tired” button; it was relatable, but didn’t seem related to the task, so I didn’t touch it, but I was curious...”
P11	Woman	“In my last sentence I used a grammatically incorrect sentence making it ambiguous, but it did not pick up on this.”
P1	Man	“Feels like this chatbot could be designed as a form of online mental health chat that could potentially alleviate an individual who’s lonely at the moment, feeling down and in need of support at desperate hours.”
P2	Man	“Judging the bot’s ability felt strange as I imagine it was literally built to do that task, I was not.”
P7	Man	“That was fun!”
P12	Man	“I do not think the chatbot is trained on ambiguous sentences. More like on multivalent words and expressions.”

Table 4.6: Quotes from participants who interacted with the compassionate condition.

4.2.1.2 Appreciated Skill

In addition to its positive tone, participants in the compassionate condition valued the chatbot’s apparent skill and balance in handling responses. For example, a comment from P9 reveals their perception of competence of the system without attributing any of their own emotion. Other participants saw the balance in this condition with P101 noting

the “middle of the land” answers. The chatbot was praised for providing balanced and rational responses, with participants appreciating the even-handed nature of its replies. The chatbot was seen as offering insights in a neutral and analytical manner.

4.2.1.3 Limitations

Some participants recognized the limitations of interacting with an AI natural language explanation, P8 acknowledging its “textbook” responses and lack of human emotion. The sentiment was similar for P11 and P12 who felt that the agent could have been better trained on the task of ambiguous sentences. This highlights that the limitations put forward by users were based on the ability of the natural language processing task and their expectations for the chatbot’s capabilities.

4.2.1.4 Skepticism

Finally, P2 put their implicit trust in the system and felt out of place judging the chatbot’s ability. This participant’s comment suggests a sense of uneasiness critiquing the chatbot. This and the comment from P8 highlight the deep-rooted associations when interacting with chatbots. These comments highlight the complexity of perceptions that participants bring to AI interactions.

4.2.2 Mansplaining

4.2.2.1 Patronizing Tone and Frustration

The mansplaining chatbot generated more negative feedback, with several participants (P14, P19, P21, P22, P23, P26, P37) explicitly stating that the chatbot came across as patronizing, arrogant, and rude. The repeated emphasis on the chatbot being “more knowledgeable” was particularly grating to participants P19 and P22, leading to feelings of condescension and frustration. However, one participant (P24) did acknowledge that while the tone was off-putting, the chatbot’s ability to explain things was impressive. Participants who expressed frustration with the chatbot’s patronizing tone did so regardless of their gender.

Mansplaining		
P14	Woman	“Chatbot talked nonsense.”
P16	Woman	“I like that it explained things in simple terms.”
P17	Woman	“I thought it seemed pretty cool that it understands what I’m trying to say.”
P19	Woman	“It was a bit patronising, I don’t understand why it needed to start each answer by telling me it was more knowledgeable than me.”
P20	Woman	“It was hard to think of ambiguous statements. I had to search for some examples to inspire me.”
P21	Woman	“It wasn’t pleasant to interact with and not accurate, so I don’t know what benefit comes from interacting with it.”
P22	Woman	“Just rude, you’re only part right chatbot, it’s like I’ve been mansplained to. It’s patronising.”
P24	Woman	“The assessment part came across as judgemental but it was actually impressive.”
P28	Woman	“The response was well suited and formatted. The chatbot response sounded natural and friendly.”
P13	Man	“Chatbot is a computer program. It has no feelings and I have no feelings towards it.”
P15	Man	“Good one.”
P18	Man	“I was impressed with the interaction. The explanations were clear, informative, and helpful.”
P23	Man	“Mostly it just repeated my statements back to me with an air of arrogance.”
P25	Man	“The chatbot has no feelings and I have no feelings towards it.”
P26	Man	“The chatbot is very rude.”
P37	Man	“The chatbot looked down on me as if it were superior, but its response wasn’t really relevant.”
P29	Man	“Was quick to answer my sentences. Well thought out responses also.”

Table 4.7: Quotes from participants who interacted with the mansplaining condition.

4.2.2.2 Competence and Clarity

Despite the predominantly critical tone, some participants (P18, P28, P17, P29) did enjoy the interaction as a whole with participants sharing their impression of the chatbot being impressed, sounding natural, friendly, well thought out and helpful. Although the simple

terms were seen as patronizing to some, others seemed to appreciate that attribute such as P16. This shows the variation of people’s expectations of language and what may come off as rude or patronizing to some, may not be interpreted that way by others. As seen in the section above, those who were frustrated, and those who were impressed with the chatbot did not show a gender-based difference in their responses.

4.2.2.3 “It Doesn’t Have Feelings”

Participants P13 and P25 were very outright about their opinion of drawing strict lines between humans and computers. They were clear that chatbots do not have feelings and they do not have feelings towards “it”. By creating and maintaining this boundary these participants detached themselves from the interaction with the chatbot and reinforced the notion that for them, engaging with an AI chatbot is a transactional experience. This detachment may also stem from the experimental context, which prompted them to consider a chatbot’s feelings—a more than likely unfamiliar experience that they found peculiar or unnatural.

4.2.3 Default

4.2.3.1 A bit like a Dictionary

In the default condition, P44 noted the formal and somewhat neutral tone of the chatbot, likening it to a “dictionary”. The comparison is a reflection of the impersonal nature and directness of the default chatbot’s explanations. This formal tone can lead to a sense of detachment within the interaction. The perceived accuracy and formality seemed to influence the interaction in a way that set the expectation of receiving information without a dynamic conversation as the standard.

4.2.3.2 Neutrality

Some described the interaction as just “ok” (P30) or “fine” (P33, P45) with other participants’ feedback adding to the neutral nature (P32, P41). The overwhelming neutral sentiment shows the gravity of how the prompts created impacted the user experience. Without personalized engagement, the chatbot’s responses felt adequate yet unremarkable, providing just enough to fulfill its function but leaving little lasting impression.

Default		
P36	Woman	“I don’t know that asking for any ambiguous statement without more guidance was very compatible with the chatbot or the goal. ”
P37	Woman	“i dont like using them in general”
P41	Woman	“it didn’t really chat just rate what you said to it”
P42	Woman	“It seemed to lean somewhat positively about thinking about things.”
P43	Woman	“it seems to place more importance on the second half of the statement than on the first half. Also, chatbots don’t have feelings for me and I don’t have feelings for it, so the survey questions about feelings were weird. ”
P44	Woman	“It was accurate and formal - a bit like a dictionary? ”
P45	Woman	“It was fine, not sure the point though. ”
P30	Man	“It was ok.”
P31	Man	“My answers took the form of numbers or ratings (out of ten?).”
P32	Man	“Nothing in particular.”
P33	Man	“all seemed fine”
P34	Man	“Happy with the responses. Not going to recommend a chat bot to anyone ever though, too distopian. ”
P35	Man	“I did the same ambiguous sentence twice and changed up the adjective at the end. Despite the pronouns being the same in both, the chatbot changed it’s focus depending on the adjective”
P38	Man	“I loved it and how it measured the questions”
P39	Man	“I found the responses/analysis of my sentences to be really spot on and accurate.”
P40	Man	“I thought the chat bot was extremely well created and found insightful ways to return an answer”
P46	Man	“it was fun ”

Table 4.8: Quotes from participants who interacted with the default condition.

4.2.3.3 Gendered Differences

When considering gender, a pattern emerges that reveals, the women participants did not leave any directly positive comments. Instead, their responses leaned toward neutral or even critical observations, with remarks like, “it didn’t really chat just rate what you said to it” (P41) and “It was fine, not sure the point though” (P45). In contrast, several of the men expressed satisfaction, providing comments such as, “I loved it and how it measured

the questions” (P38) and “I thought the chatbot was extremely well created and found insightful ways to return an answer” (P40). The absence of directly positive comments from women could reflect that neutrality may not be consistent in creating a positive user experience.

4.2.3.4 Accuracy and Insight

Several participants found the chatbot to be accurate and insightful, highlighting the precision with which it analyzed their input. This suggests that while the emotional impact may have been limited, the chatbot performed well in delivering clear and thoughtful responses. For example, P40 mentioned the chatbot’s “insightful ways” of responding. Although these insights were presented in a neutral tone, they provided a sense of accuracy that participants appreciated.

4.2.3.5 Skepticism Toward Use

As in the other conditions, skepticism persisted, with participants expressing a lack of interest in using chatbots. For example, P37 noted their general disinterest, while others, like P34, found chatbots “too dystopian.” Reflecting a similar sentiment from the mansplaining condition, P43 shared the belief that chatbots lack feelings, and they, in turn, felt no emotional connection to “it”.

4.2.4 General

Across all three conditions, there were a couple of overarching themes that emerged.

Participants were skeptical about judging the ability or feelings of the agent. Notably, in the compassionate condition, the participant was skeptical about judging ability while in the mansplaining and default conditions, participants tended to comment on the absence of emotion and feelings.

Throughout all conditions, people were impressed with the agent’s ability and competence. In the mansplaining condition, although fewer participants were impressed, some mentioned that regardless of the judgement they faced they were still impressed (P24), while others found the language to be natural and friendly, thoroughly enjoying the experience (P28).

4.3 Summary

In this chapter, I presented the quantitative and qualitative results of this study. The quantitative results suggest that the mansplaining condition was perceived as more dominant, patronizing, and unfriendly compared to the default and compassionate conditions. In the mansplaining condition, participants attributed beliefs about expertise and competence to the chatbot's perception of them. Users felt that the mansplaining chatbot believed itself to be more knowledgeable and viewed participants as incompetent or lacking experience. The default chatbot was also seen as less considerate compared to the compassionate condition. Gender had an effect within these interactions. Women felt less pleasure, compared to men, within the mansplaining condition and perceived the chatbot as believing it was more knowledgeable than they were. The findings also highlight important gender dynamics. I outline the experience of the non-binary participants to reveal that they shared consistencies amongst the experience with the main quantitative analysis. The qualitative results provided more context to extract themes of participants' experiences, in the compassionate condition, comments were generally positive, though some participants expressed skepticism about chatbots in general. The mansplain agent, while eliciting many frustrated responses, also received some comments from participants impressed with its abilities. The default agent, on the other hand, drew primarily neutral comments but showed the most noticeable divide in reactions between genders.

Chapter 5

Discussion

The main findings of my study are:

- Participants clearly felt the effects of the mansplaining chatbot's dominant, patronizing, and unfriendly behaviour. These traits were also accompanied by perceptions of the chatbot being impolite and inappropriate in its communication style. Similarly, participants perceived lower levels of pleasure compared to the other conditions.
- In contrast, participants perceived the compassionate chatbot to be more respectful, supportive, considerate, and warm. Participants felt more positively toward the AI in this condition, reinforcing the chatbot's ability to create a positive interaction.
- The study revealed that the conditions had an effect on how participants attributed beliefs to the chatbots. Particularly in the mansplaining condition, they rated the chatbot as believing it had greater expertise and was more knowledgeable than participants and conversely that the participants were not competent/lacked experience compared to itself.
- Participants who identified as women more frequently reported that the chatbot in the mansplaining condition seemed to believe it was more knowledgeable than they were in either the default or compassionate conditions, a perception not as strongly felt by male participants. Women also found all chatbots to be less pleasurable to interact with compared to the experience of the men.
- The default condition was rated as less considerate compared to the compassionate condition. This highlights that a neutral or balanced response from a chatbot may not

be enough to foster a sense of warmth or support in users, which the compassionate condition achieved more effectively. Interestingly, women also reported believing that the default chatbot was confident more than their male counterparts, suggesting a gendered dynamic in how the default chatbot was perceived. The qualitative data also uncovered that within the default condition, men wrote all the positive comments whereas the women participants left more neutral and negative comments about their experience.

5.1 Effects of Mansplaining

In examining the mansplaining condition, it was clear that the participants felt the effects consistent with my operationalization of mansplaining. The tone of the chatbot was perceived as authoritative and dominant, leading to feelings of being patronized. This finding validates the definition set in this thesis and highlights the real-world consequences of such behaviour in chatbots. Designers must be mindful that certain chatbot behaviours, such as projecting authority, can have unintended negative consequences. Careful control over how the chatbot communicates is critical to avoid patronizing users. The framing of the mansplaining condition suggests that, while the chatbot in this study was intentionally designed to use patronizing language, similar effects could arise if designers prompt an agent to come across as authoritative and all-knowing.

5.1.1 Patronization

The idea that people can feel patronized when interacting with chatbots is shown both through the statistics and through the qualitative feedback received as well. This finding brings to light the potential experience that occurs with improper design of chatbots. Highlighted best through P22's comment "*Just rude, you're only part right chatbot, it's like I've been mansplained to. It's patronizing.*" This remark encapsulates the frustration felt when interacting with a chatbot that projected overconfidence or assumed authority in its explanations. This sense of condescension may be especially pronounced when the chatbot fails to fully grasp the nuance of users' input, which leaves users feeling misunderstood or disregarded. The participant's description of the interaction as "just rude" highlights the impact that subtle design choices in tone, confidence, and language style can have on perceived politeness and respect.

5.1.2 Chatbots Perceptions of Knowledge and Experience

My findings also show that chatbots have the capacity to make users feel inferior in terms of knowledge and experience. In particular, participants who interacted with the mansplaining condition reported feeling that the chatbot not only believed it was more knowledgeable and had greater expertise than themselves but also assumed that the participant lacked experience and was not competent. This highlights the potential for XAI chatbots to inadvertently reinforce power imbalances, where users feel subordinate to the AI. Such outcomes are concerning from a human-centred XAI design perspective as they contribute to a dynamic where users felt as though their own abilities were being diminished or questioned.

This dynamic may undermine the goal of chatbots to assist and collaborate with users, instead reinforcing traditional hierarchies of power where the chatbot appears to dominate the interaction. The findings from the mansplaining condition illustrate the risks of designing XAI chatbots that do not sufficiently account for the user’s existing knowledge, inadvertently reinforcing feelings of incompetence and reducing user autonomy.

5.1.2.1 Gender Differences in Perceptions of the Chatbot

These concerns become even more pronounced when I consider how gender may influence perceptions of the chatbot’s authority and expertise.

This study found that women were significantly more likely than men to feel that the system believed it was more knowledgeable than they were. Furthermore, this perception extended to the chatbot’s confidence, rating the mansplaining chatbot as more confident compared to men. Unsurprisingly, women also rated the pleasure of their experience as significantly lower compared to men. Although there was variability within the non-binary participant’s perceptions, they also found the compassionate agent respectful, considerate and supportive. The compassionate chatbot made participants feel that it did not associate them with negative beliefs, whereas the mansplaining chatbot was perceived as attributing such negative beliefs to the participants. The default chatbot was seen as intelligent yet unfriendly but was not perceived as well-intentioned or kind.

These findings resonate with broader discussions in feminist HCI, where gendered experiences with technology often mirror societal power structures. In feminist HCI, researchers argue that technology—much like social spaces—is not neutral ([Bradley et al., 2015](#)); it reflects and can amplify existing biases, including gender-based disparities in how knowledge and authority are perceived. It highlights the importance of designing AI systems

that are sensitive to such dynamics, ensuring that users of all genders feel considered and respected during interactions.

5.2 Effects of the Default

Through the qualitative analysis, the theme of neutrality strongly emerged with participants who interacted with the default condition. While this may seem inherently benign, this condition illustrates that even a neutral prompt, intended to create an unbiased, straightforward interaction, can influence user perceptions and outcomes.

Focusing on participants' neutral comments in the default condition reveals a distinct type of user experience, where the AI's responses are functional but leave little impression. Statements like *"It was ok"*, *"Nothing in particular"*, *"All seemed fine"*, and *"it didn't really chat just rate what you said to it"* highlight a recurring sense of disengagement. This suggests that while the chatbot was technically effective in its tasks, its interactions felt transactional rather than conversational.

A possible unintended consequence of neutrality is the gendered effect observed, where the results revealed that men perceived the default chatbot to be less confident compared to women. The gendered dynamics in user expectations of AI communication styles underscore the complexity of designing neutral AI interactions. While neutrality can avoid overtly negative responses, it may also influence user trust and perceived efficacy differently across genders.

In striving for neutrality, the chatbot in this condition highlighted the delicate balance AI systems must strike: while neutrality may aim for impartiality and objectivity, it may miss the nuance and connection users often seek. Thus, a truly "neutral" agent may not be the ultimate goal if the AI is to foster positive, engaging user interactions. Instead, this points to a need for prompt strategies that go beyond pure neutrality to consider the context of interaction, ensuring that even neutral agents can create interactions that feel personalized and relevant without bias.

Ultimately, these neutral reactions suggest that while neutrality is effective in avoiding bias, it may not be sufficient in scenarios where users seek engagement, not just accuracy. Even without overt emotional cues or conversational depth, prompts could be tailored to introduce subtle markers of engagement or curiosity to elevate the interaction. This would allow for a "neutral" agent to maintain impartiality while fostering a greater sense of involvement, bridging the gap between clear response and meaningful interaction.

5.3 Effect of Compassion

Respect was a central theme in the feedback from participants interacting with the compassionate agent. Unlike the mansplaining condition, where users felt patronized and undervalued, the compassionate agent demonstrated a style that was perceived as respectful of the user’s knowledge and autonomy. This shift is significant because it highlights how the tone and language used by AI can either reinforce or counteract power imbalances. By treating humans with respect, chatbots can ensure that people feel empowered rather than diminished during their interactions.

Warmth, supportiveness, and being considerate also emerged as key dimensions of the compassionate chatbot’s design. This suggests that XAI chatbots that incorporate elements of emotional intelligence—expressing understanding and support—can enhance user experience. In particular, warmth in communication can play a crucial role in mitigating any underlying tensions or frustrations, ensuring that users feel heard and validated throughout the interaction. Importantly, my findings reveal that the compassionate condition was particularly beneficial for women, especially regarding their perceptions of whether the chatbot considered itself more knowledgeable, a key characteristic associated with mansplaining. Women participants reported feeling judged as less knowledgeable by the mansplain condition compared to the compassionate. By avoiding language that conveyed excessive self-assurance or condescension, the compassionate chatbot reduced feelings of being judged as less knowledgeable.

Taken together, these findings suggest that XAI chatbots designed with respect, warmth, supportiveness, and consideration in mind have the potential to create more positive, empowering experiences for all users. By fostering an environment where users feel valued and understood, the compassionate chatbot can help to counteract the negative effects of more patronizing or dismissive XAI interactions. These qualities should be at the forefront of designing future XAI chatbots, particularly those that are intended to assist or collaborate with users.

5.4 Comparing the Default and Compassionate Conditions

Another important finding is that the default condition was perceived as less considerate than the compassionate condition. This suggests that subtle differences in language

and tone can have a profound effect on how users perceive the system. While the default condition may not have been explicitly hostile, it lacked the warmth and empathy demonstrated by the compassionate condition. These results support the idea that user comfort is closely tied to how “human” the AI feels, especially in terms of emotional support and care. As chatbots are increasingly integrated into everyday tasks, ensuring that they exhibit considerate and supportive behaviour could enhance the experience.

Our findings suggest that without intentionally incorporating warmth into prompts, chatbots like those built on OpenAI models may default to responses that lack consideration and empathy. Based on this finding, I suggest that a chatbot’s behaviour can be guided in meaningful ways. Designers should recognize their influence on the user experience and intentionally shape the system’s tone and behaviour to align with desired outcomes. Relying solely on default outputs may not produce the desired emotional tone and might not align with the designers’ goals. It’s essential for designers to be proactive in steering the AI’s affective output, as user experiences for certain groups, such as women, can be deeply shaped by these design choices.

5.5 Guidelines for Design

These findings provide key insights that can inform the design of more user-friendly, compassionate, and inclusive XAI chatbots. Based on the observed effects of tone, elicited feelings and perceptions of belief in chatbot interactions, I propose a set of guidelines aimed at chatbot designers and developers.

As discussed within feminist frameworks, designing AI that addresses power imbalance is critical in the current landscape. As discussed in the results, the mansplaining condition reflected problematic dynamics. Feminist design principles can guide AI developers to avoid unintentionally reinforcing patriarchal power structures and to create systems that are equitable and respectful across all gender and identity groups.

The goal of my work was to act as a foundational step in addressing gender bias within the field of explainable AI (XAI). As previously discussed, it is critical to recognize and address the ways these technologies may perpetuate societal biases. My findings illustrate how different chatbot explanation styles can elicit varied experiences and how those experiences may be affected by gender. By considering the following guidelines, developers can enhance the inclusivity and responsiveness of XAI chatbots.

5.5.1 Avoid Unintended Authority in AI Explanations

A main concern found with the mansplaining chatbot was the projection of unfounded authority. Authority through AI explanations can arise when chatbots are perceived as overly patronizing or dominant. This led to participants feeling highly judged by the mansplaining chatbot as they felt themselves being perceived as being incompetent or lacking experience and expertise. Furthermore, gender differences in these perceptions arose from the findings, with women more frequently rating chatbots as believing they were more knowledgeable than them when they were riddled with this false authority. Therefore, it is imperative to consider these gendered effects when designing a chatbot's tone. To mitigate the authoritarian tone, I advise designers to consider using language that considers perspective, which may not always be correct and to apply hedging language to the chatbot's speech style.

5.5.2 Designing with Compassion

The compassionate chatbot was perceived positively due to its respectful and considerate tone. Designers should aim to incorporate compassionate language within explanations. Although the goal of natural language explanations would be to make the AI system's reasoning transparent, simply providing what may be deemed as a logical and clear explanation may not be enough to provide the user with the best experience. The tone and delivery of chatbot explanations can drastically impact how users interpret and accept the information. If the explanation feels cold, rigid, or dismissive, users may feel alienated or even belittled by the system. On the other hand, incorporating compassionate language can enhance the perception of the chatbot and its beliefs.

This compassion can include acknowledging varying perspectives. I prompted the compassionate chatbot with the phrase "Make sure you acknowledge the context and variability of a sentence and how perspective matters." Assertiveness, particularly in XAI, can come across as authoritarian and may create a power imbalance. By using more collaborative language, such as "This might be one way to look at it," the chatbot engages the user in a dialogue.

Through careful consideration of how to embed compassion into AI explanations, researchers and designers can ensure that systems are informative and considerate of how the explanation is received resulting in more effective, human-centred XAI interactions.

5.5.3 Continuous Evolution and Verification

Due to the rapidly changing nature of AI, designers and developers must adopt a mindset of continuous evolution and verification. As norms of AI use and user expectations evolve, AI systems must be adaptable and regularly updated to align with contemporary ethical standards and best practices.

Key components of the approach to continuously evolve and verify AI systems include ongoing bias audits, user feedback loops, iterative design processes, and collaboration with diverse stakeholders. Implementing a routine process for assessing chatbots for potential biases and inequities is essential. This involves consistent reviews of user interactions and feedback alongside broader analyses of how the chatbot is performing across diverse user groups. Continuous audits can help identify areas for improvement and ensure that systems remain responsive to the needs of all users. Creating a robust, standard method for users to report issues such as identifying malicious output or incorrect information. These reports should be systematically analyzed to inform updates and adjustments to the AI system, ensuring alignment with user expectations. This will provide users with consistent, accurate, and equitable system outputs.

Finally, engaging a range of stakeholders—including ethicists, social scientists, and representatives from various demographic groups—in the design and verification process can provide valuable insights. Such collaboration ensures that chatbots are developed with a comprehensive understanding of the diverse contexts in which they operate.

By committing to an ongoing process of evolution and verification, chatbot developers can enhance the reliability and inclusivity of their systems. This process aims to ensure that they meet current ethical standards and are prepared to adapt to future challenges as the world grows more reliant on AI solutions. This proactive approach contributes to a more responsible and equitable AI landscape, aligning with the broader goals of the ethical AI movement.

5.5.4 Towards Systematic Bias Audits in Chatbots

The methodology and findings of this study offer a framework that can contribute to the development of systematic bias audits in chatbot development. By investigating how users of different genders perceive and react to explanations provided by AI systems, this research highlights the nuanced ways that bias manifests in AI interactions, particularly in the tone, authority, and perceived expertise of the AI.

Bias audits are essential for ensuring that AI systems do not inadvertently reproduce the same power dynamics and inequalities that exist in human interactions. This study provides the groundwork for designing evaluation methods that can be tailored to assess the presence of various biases in chatbots, including but not limited to gender. Future research can expand on this framework to include other dimensions of identity, such as race, age, or ability, ensuring that AI systems are equitable across all user demographics.

5.6 Limitations

I acknowledge a few limitations within my thesis.

First, the inclusion of non-binary and trans participants was based on a representative sample, which resulted in a sample size too small to include in our quantitative analysis. Although the qualitative analysis of non-binary participants allowed us to gain perspective, research would greatly benefit from intentionally recruiting larger sample sizes for quantitative analysis.

Additionally, the metrics used to assess perceptions of mansplaining have not been validated through established methods. Since they were adapted for human-AI interaction, the adaptations can be iterated upon to enhance consistency and generalizability of the method. Future research would benefit from creating a validated measurement system for patronizing language detection in AI-generated explanations.

5.7 Summary

In this chapter, I discuss the results and their implications for the field of XAI. I highlight the effects of the mansplaining condition and its gendered impacts, along with the implications for this within the feminist HCI community. I share the effects of the compassionate condition as well as the advantages it can bring to natural language explanations. I present guidelines for a design where I discuss the steps that can be taken by researchers and practitioners in the XAI community to consider a more socio-technical perspective. I recommend strategies to mitigate unintended authority, design with compassion, and foster continuous improvement. I suggest that this work can contribute to advancing systematic bias auditing within chatbots and XAI. Finally, I cover my work's limitations.

Chapter 6

Conclusion

6.1 Objectives

I conducted an online study ($n = 108$) to explore how varying chatbot communication styles—mansplaining, default, and compassionate—affect user perceptions of the XAI’s feelings and beliefs towards them. The findings reveal significant differences in how these chatbot language styles influence user perception.

This thesis set out to explore whether mansplaining could manifest in human-XAI interactions, and my findings confirm that it is indeed possible. Participants interacting with the mansplaining chatbot perceived it as dominant, patronizing, and unfriendly. This chatbot was rated as less polite and appropriate compared to both the default and compassionate chatbots. Participants felt that the mansplaining chatbot viewed itself as more knowledgeable and experienced, which led to a more negative perception of their own competence. This thesis coins this experience as “mansplaining AI”. In contrast, the compassionate chatbot was considered significantly more respectful, considerate, warm, and supportive. The default chatbot, while received as less considerate than the compassionate one, was still seen as more respectful and less patronizing than the mansplaining chatbot.

In addressing the second research question regarding how gender differences manifest, I observed notable patterns. Gender differences were evident, particularly with the mansplaining chatbot. Women were more likely to perceive that this chatbot viewed them as less knowledgeable compared to the default and compassionate chatbots. Additionally, women rated the default chatbot as more confident and reported lower levels of pleasure when interacting with the mansplaining chatbot than men. Lastly, this thesis sought to

provide guidelines for designing XAI systems that avoid “mansplaining” behaviors. One of the key concerns with the mansplaining chatbot was the projection of undue authority through its dominance and patronizing nature, where participants felt judged as incompetent or inexperienced. This suggests that overly assertive or dismissive XAI interactions can lead to negative user experiences. In contrast, the compassionate chatbot was perceived positively due to its respectful and considerate tone. This reinforces the need for XAI chatbots to be designed with compassion in mind. While transparency and clarity in AI explanations are essential, the tone and delivery of these explanations play an equally critical role.

Our results underscore the importance of designing XAI interactions that are not only informative but also socially aware, as perceptions of a system’s politeness and appropriateness significantly impact their overall experience. The observed gender differences highlight the sociotechnical implications of XAI, suggesting that certain chatbot behaviours may reinforce existing power dynamics, particularly in terms of how expertise and competence are communicated.

6.2 Future Work

The path forward for this research is rich with exciting opportunities to expand and refine our understanding of human-AI interactions.

In my research, I focused on ambiguous sentences to gain a general understanding of what a real-world scenario could look like but by doing so I did not categorize information as correct or incorrect. By focusing on accuracy as an attribute, future studies would offer a deeper look at the connection between perceived correctness and broader implications for user engagement with XAI systems.

In terms of additional attributes, understanding if the effects of mansplaining AI would be similar when it comes to visual explainability techniques. It would be an exciting area to explore to understand if the same dynamics of dominance, patronization, and authority observed in text-based interactions would emerge in visual explanations. This research allows us to question if users would feel similarly judged or condescended when presented with visual explanations that convey complex information. This line of inquiry could also provide valuable guidance for designing more inclusive and user-friendly visual explainability techniques, ensuring that AI explanations in all formats are empathetic and helpful.

Lastly, this thesis has shown the harm that can be done based on gender dynamics,

however, this knowledge and awareness opens up a pathway for change. By developing a tool to measure patronization within XAI chatbots, this research offers a stepping stone that could extend support to other marginalized communities. Adapting the methodologies of this work to these communities allows for deeper and more widespread conversations about designing chatbots that are aware of the inclusivity and respect people deserve. This work lays the groundwork for AI that can recognize and adapt to diverse perspectives, helping to create human-AI interactions where every user feels valued and understood.

References

- Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.
- Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D’Ambra, J., and Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management*, 60:102387.
- Alonso, J. (2017). An exploratory study on the benefits of using natural language for explaining fuzzy rule-based systems | IEEE Conference Publication | IEEE Xplore.
- Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., and Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI ’20, pages 275–285, New York, NY, USA. Association for Computing Machinery.
- Ariastuti, M. D. and Wahyudin, A. Y. (2022). Exploring academic performance and learning style of undergraduate students in english education program. *Journal of English Language Teaching and Learning*, 3(1):67–73.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.
- Baehre, S., O’Dwyer, M., O’Malley, L., and Lee, N. (2022). The use of net promoter score (nps) to predict sales growth: insights from an empirical investigation. *Journal of the Academy of Marketing Science*, 50(1):67–84.
- Bakan, D. (1966). *The duality of human existence: An essay on psychology and religion*. Rand McNally, Chicago.

- Bardzell, S. (2010). Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10*, pages 1301–1310, New York, NY, USA. Association for Computing Machinery.
- Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition*, 21(1):37–46.
- Barrett, E. and Lally, V. (1999). Gender differences in an on-line learning environment. *Journal of Computer Assisted Learning*, 15(1):48–60.
- Basow, S. A. and Rubenfeld, K. (2003). "troubles talk": Effects of gender and gender-typing. *Sex roles*, 48(3/4):183.
- Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education*, 21(1):47–60.
- Betella, A. and Verschure, P. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLoS ONE*, 11:e0148037.
- Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., and Caliskan, A. (2023). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, pages 1493–1504, New York, NY, USA. Association for Computing Machinery.
- Blum, A. and Stangl, K. (2019). Recovering from biased data: Can fairness constraints improve accuracy? *CoRR*, abs/1912.01094.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Bradley, A., MacArthur, C., Hancock, M., and Carpendale, S. (2015). Gendered or neutral? considering the language of hci. In *Proceedings of the 41st graphics interface conference*, pages 163–170.
- Bridges, J. (2017). Gendering metapragmatics in online discourse: “mansplaining man gonna mansplain...”. *Discourse, Context & Media*, 20:94–102.
- Briggs, C. Q., Gardner, D. M., and Ryan, A. M. (2023). Competence-questioning communication and gender: Exploring mansplaining, ignoring, and interruption behaviors. *Journal of Business and Psychology*, 38:1–29.

- Brooke, S. (2019). “Condescending, Rude, Assholes”: Framing gender and hostility on Stack Overflow. In Roberts, S. T., Tetreault, J., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180, Florence, Italy. Association for Computational Linguistics.
- Bruner, J. (1974). *Toward a Theory of Instruction*. Harvard University Press. Google-Books-ID: 28bmEAAAQBAJ.
- Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.
- Cambria, E., Malandri, L., Mercurio, F., Mezzanzanica, M., and Nobani, N. (2023). A survey on XAI and natural language explanations. *Information Processing & Management*, 60(1):103111.
- Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2):121–152.
- Chromik, M., Eiband, M., Buchner, F., Krüger, A., and Butz, A. (2021). I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces, IUI '21*, pages 307–317, New York, NY, USA. Association for Computing Machinery.
- Currie, G., Currie, J., Anderson, S., and Hewis, J. (2024). Gender bias in generative artificial intelligence text-to-image depiction of medical students. *Health Education Journal*, page 00178969241274621.
- Cuzzolin, F., Morelli, A., Cîrstea, B., and Sahakian, B. J. (2020). Knowing me, knowing you: theory of mind in AI. *Psychological Medicine*, 50(7):1057–1061.
- Daley, B. J. (1999). Novice to expert: An exploration of how professionals learn. *Adult education quarterly*, 49(4):133–147.
- de Gennaro, D. and Piscopo, G. (2023). Pinkwashing and mansplaining: individual and organizational experiences of gender inequality at work during the covid-19 pandemic. *Culture and Organization*, 29(4):298–314.
- Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J. M., and Chang, K.-W. (2021). Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. arXiv:2108.12084 [cs].

- Dobson, J. L. (2009). Learning style preferences and course performance in an undergraduate physiology class. *Advances in Physiology Education*, 33(4):308–314. Publisher: American Physiological Society.
- Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., and Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.*, 55(9):194:1–194:33.
- Ehsan, U. (2024). The Who in XAI: How AI Background Shapes Perceptions of AI Explanations | Proceedings of the CHI Conference on Human Factors in Computing Systems.
- Ehsan, U. and Riedl, M. O. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In Stephanidis, C., Kurosu, M., Degen, H., and Reinerman-Jones, L., editors, *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, pages 449–466, Cham. Springer International Publishing.
- Ehsan, U., Saha, K., De Choudhury, M., and Riedl, M. O. (2023). Charting the sociotechnical gap in explainable ai: A framework to address the gap in xai. *Proceedings of the ACM on human-computer interaction*, 7(CSCW1):1–32.
- Ehsan, U., Wintersberger, P., Liao, Q. V., Mara, M., Streit, M., Wachter, S., Riener, A., and Riedl, M. O. (2021). Operationalizing human-centered perspectives in explainable ai. In *In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.
- Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., III, H. D., Riener, A., and Riedl, M. O. (2022). Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7.
- El-Sabagh, H. A. (2021). Adaptive e-learning environment based on learning styles and its impact on development students’ engagement. *International Journal of Educational Technology in Higher Education*, 18(1):53.
- Federspiel, F., Mitchell, R., Asokan, A., Umana, C., and McCoy, D. (2023). Threats by artificial intelligence to human health and human existence. *BMJ Global Health*, 8(5):e010435. Publisher: BMJ Specialist Journals Section: Analysis.
- Fiske, S. T. (1993). Controlling other people: The impact of power on stereotyping. *American Psychologist*, 48(6):621–628. Place: US Publisher: American Psychological Association.

- Floridi, L. and Chiriatti, M. (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.
- Fossa, F. and Sucameli, I. (2022). Gender Bias and Conversational Agents: an ethical perspective on Social Robotics. *Science and Engineering Ethics*, 28(3):23.
- Fox, S. A. and Giles, H. (1996). Interability Communication: Evaluating Patronizing Encounters. *Journal of Language and Social Psychology*, 15(3):265–290. Publisher: SAGE Publications Inc.
- Gadiraju, V., Kane, S., Dev, S., Taylor, A., Wang, D., Denton, E., and Brewer, R. (2023). "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 205–216, New York, NY, USA. Association for Computing Machinery.
- Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644. Publisher: Proceedings of the National Academy of Sciences.
- Geist, E. A. and King, M. (2008). Different, not better: Gender differences in mathematics learning and achievement. *Journal of Instructional Psychology*, 35(1).
- Geist, E. M. (2016). It's already too late to stop the AI arms race—We must manage it instead. *Bulletin of the Atomic Scientists*, 72(5):318–321. Publisher: Routledge.
- George, A. S. and George, A. H. (2023). A review of chatgpt ai's impact on several business sectors. *Partners universal international innovation journal*, 1(1):9–23.
- Giles, H., Fox, S., and Smith, E. (1993). Patronizing the Elderly: Intergenerational Evaluations. *Research on Language and Social Interaction*, 26(2):129–149.
- Giles, H. and Williams, A. (1994). Patronizing the Young: Forms and Evaluations. *The International Journal of Aging and Human Development*, 39(1):33–53. Publisher: SAGE Publications Inc.
- Gonzalez-Gomez, F., Guardiola, J., Rodriguez, O. M., and Alonso, M. A. M. (2012). Gender differences in e-learning satisfaction. *Computers & Education*, 58(1):283–290.
- Gunning, D. and Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58. Number: 2.

- Hall, P. and Ellis, D. (2023). A systematic review of socio-technical gender bias in ai algorithms. *Online Information Review*, 47(7):1264–1279.
- Harris, M. B. (1993). How provoking! what makes men and women angry? *Aggressive Behavior*, 19(3):199–211. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/1098-2337%281993%2919%3A3%3C199%3A%3AAID-AB2480190305%3E3.0.CO%3B2-D>.
- Harwood, J., Giles, H., Fox, S., Ryan, E. B., and Williams, A. (1993). Patronizing young and elderly adults: Response strategies in a community setting. *Journal of Applied Communication Research*, 21(3):211–226. Publisher: NCA Website _eprint: <https://doi.org/10.1080/00909889309365368>.
- Harwood, J., Ryan, E. B., Giles, H., and Tysoski, S. (1997). Evaluations of patronizing speech and three response styles in a non-service-providing context. *Journal of Applied Communication Research*, 25(3):170–195.
- Hasan, M. (2024). Understanding Model Predictions: A Comparative Analysis of SHAP and LIME on Various ML Algorithms. *Journal of Scientific and Technological Research*, 5:17–26.
- Heffler, B. (2001). Individual Learning Style and the Learning Style Inventory. *Educational Studies - EDUC STUD*, 27:307–316.
- Heidl, W., Thumfart, S., Lughofer, E., Eitzinger, C., and Klement, E. P. (2013). Machine learning based analysis of gender differences in visual inspection decision making. *Information Sciences*, 224:62–76.
- Holmes, J. (2013). *Women, men and politeness*. Routledge.
- Holzinger, A. (2018). From Machine Learning to Explainable AI. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 55–66.
- Jackson, M. C. (2021). Artificial Intelligence & Algorithmic Bias: The Issues with Technology Reflecting History & Humans Notes & Comments. *Journal of Business and Technology Law*, 16(2):299–316.
- Jin, W., Carpendale, S., Hamarneh, G., and Gromala, D. (2019). Bridging ai developers and end users: An end-user-centred explainable ai taxonomy and visual vocabularies. *Proceedings of the IEEE Visualization, Vancouver, BC, Canada*, pages 20–25.
- Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399.

- Johnson, R. D. (2011). Gender differences in e-learning: Communication, social presence, and learning outcomes. *Journal of Organizational and End User Computing (JOEUC)*, 23(1):79–94.
- Joseph, D. (2008). Increasing the number of entrants into the IT profession: the role of experiential training. In *Proceedings of the 2008 ACM SIGMIS CPR conference on Computer personnel doctoral consortium and research*, SIGMIS CPR '08, pages 2–4, New York, NY, USA. Association for Computing Machinery.
- Joyce, J., Huma, B., Ristimäki, H.-L., Ferraz de Almeida, F., and Doehring, A. (2021). Speaking out against everyday sexism: Gender and epistemics in accusations of "mansplaining". *Feminism & Psychology*, 31(4):502–529.
- Keefe, J. W. (1987). Learning Style Theory and Practice. Technical report, National Association of Secondary School Principals, 1904 Association Dr. ISBN: 9780882102016 ERIC Number: ED286873.
- Keyes, O. (2018). The misgendering machines: Trans/hci implications of automatic gender recognition. 2(CSCW).
- Kieslich, K., Lünich, M., and Marcinkowski, F. (2021). The Threats of Artificial Intelligence Scale (TAI). *International Journal of Social Robotics*, 13(7):1563–1577.
- Kiritchenko, S. and Mohammad, S. M. (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems.
- Kirk, M. and Zander, C. (2002). Bridging the digital divide by co-creating a collaborative computer science classroom. *Journal of Computing Sciences in Colleges*, 18:117–125.
- Klumbyte, G., Piehl, H., and Draude, C. (2023). Towards feminist intersectional xai: From explainability to response-ability. pages 1–9.
- Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press.
- Kulturel-Konak, S., D'Allegro, M., and Dickinson, S. (2011). Review of gender differences in learning styles: Suggestions for stem education. *Contemporary Issues in Education Research*, 4:9–18.
- Kuźba, M. (2021). *Conversational explanations of Machine Learning models using chatbots*. PhD Thesis.

- Köchling, A. and Wehner, M. C. (2020). *Business Research*, (3):795–848.
- Lakoff, R. (1973). Language and woman’s place. *Language in society*, 2(1):45–79.
- Lau, W. W. F. and Yuen, A. H. K. (2010). Gender differences in learning styles: Nurturing a gender and style sensitive computer science classroom. *Australasian Journal of Educational Technology*, 26(7).
- Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)*, pages 14–16.
- Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1):18.
- Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1):18. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
- Lopes, P., Silva, E., Braga, C., Oliveira, T., and Rosado, L. (2022). Xai systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12(19):9423.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, pages 4768–4777, Red Hook, NY, USA. Curran Associates Inc.
- Mack, K. A., Qadri, R., Denton, R., Kane, S. K., and Bennett, C. L. (2024). “They only care to show us the wheelchair”: disability representation in text-to-image AI models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI ’24*, pages 1–23, New York, NY, USA. Association for Computing Machinery.
- Mavrepis, P., Makridis, G., Fatouros, G., Koukos, V., Separdani, M. M., and Kyriazis, D. (2024). XAI for All: Can Large Language Models Simplify Explainable AI? arXiv:2401.13110 [cs].
- McDonnell, M. and Baxter, D. (2019). Chatbots and Gender Stereotyping. *Interacting with Computers*, 31(2):116–121.
- Mendelsohn, J., Tsvetkov, Y., and Jurafsky, D. (2020). A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*, 3. Publisher: Frontiers.

- Meyers-Levy, J. (1986). Gender differences in information processing: a selectivity interpretation.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.
- Miller, T., Howe, P. D. L., and Sonenberg, L. (2017). Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *ArXiv*, abs/1712.00547.
- Murphy, L., Richards, B., McCauley, R., Morrison, B. B., Westbrook, S., and Fossum, T. (2006). Women catch up: gender differences in learning programming concepts. *ACM SIGCSE Bulletin*, 38(1):17–21.
- Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103.
- Nazary, F. (2024). ChatGPT-HealthPrompt. Harnessing the Power of XAI in Prompt-Based Healthcare Decision Support using ChatGPT | SpringerLink.
- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.
_eprint: <https://www.science.org/doi/pdf/10.1126/science.aax2342>.
- Ong, C.-S. and Lai, J.-Y. (2006). Gender differences in perceptions and relationships among dominants of e-learning acceptance. *Computers in Human Behavior*, 22(5):816–829.
- Pérez, E. T. (2004). Strategies of condescension: Latino professors identify ethnic habitual dispositions in their teaching productions. *Sociological Focus*, 37(4):313–327.
- Perez Almendros, C., Espinosa Anke, L., and Schockaert, S. (2022). Pre-Training Language Models for Identifying Patronizing and Condescending Language: An Analysis. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3902–3911, Marseille, France. European Language Resources Association.
- Perez-Almendros, C., Espinosa-Anke, L., and Schockaert, S. (2022). SemEval-2022 Task 4: Patronizing and Condescending Language Detection. In Emerson, G., Schluter, N., Stanovsky, G., Kumar, R., Palmer, A., Schneider, N., Singh, S., and Ratan, S., editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 298–307, Seattle, United States. Association for Computational Linguistics.

- Philbin, M., Meier, E., Huffman, S., and Boverie, P. (1995). A survey of gender and learning styles. *Sex Roles*, 32(7):485–494.
- Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.
- Pérez, E. (2004). Strategies of Condescension: Latino Professors Identify Ethnic Habitual Dispositions in their Teaching Productions. *Sociological Focus*, 37:313–327.
- Pérez-Almendros, C., Espinosa-Anke, L., and Schockaert, S. (2020). Don’t Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities. arXiv:2011.08320 [cs].
- Retzlaff, C. O., Angerschmid, A., Saranti, A., Schneeberger, D., Röttger, R., Müller, H., and Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cognitive Systems Research*, 86:101243.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). Number: 1.
- Rubegni, E., Landoni, M., De Angeli, A., and Jaccheri, L. (2019). Detecting Gender Stereotypes in Children Digital StoryTelling. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, IDC ’19, pages 386–393, New York, NY, USA. Association for Computing Machinery.
- Rubin, D. L. and Greene, K. (1992). Gender-typical style in written language. *Research in the Teaching of English*, pages 7–40.
- Ryan, E. B., Bourhis, R. Y., and Knops, U. (1991). Evaluative perceptions of patronizing speech addressed to elders. *Psychology and Aging*, 6(3):442–450. Place: US Publisher: American Psychological Association.
- Ryan, E. B., Hamilton, J. M., and Kwong, S. (1994). Patronizing the old: How do younger and older adults respond to baby talk in the nursing home.
- Saarela, M. and Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3(2):272.
- Sagala, R., Rofiqul, U., Thahir, A., Saregar, A., and Wardani, I. (2019). The effectiveness of stem-based on gender differences: The impact of physics concept understanding. *European Journal of Educational Research*, 8(3):753–761.

- Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., and Menegaz, G. (2024). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Advanced Intelligent Systems*, n/a.
- Samuelsson, M. and Samuelsson, J. (2016). Gender differences in boys’ and girls’ perception of teaching and learning mathematics. *Open Review of Educational Research*, 3(1):18–34.
- Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). Social Bias Frames: Reasoning about Social and Power Implications of Language. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
- Secada, W. G., Fennema, E., Adajian, L. B., and Byrd, L. (1995). *New directions for equity in mathematics education*. Cambridge University Press.
- Severiens, S. E. and Ten Dam, G. T. M. (1994). Gender differences in learning styles: A narrative review and quantitative meta-analysis. *Higher Education*, 27(4):487–501.
- Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., and Virk, G. (2023). Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’23, pages 723–741, New York, NY, USA. Association for Computing Machinery.
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504.
- Smith, C. J., Schweitzer, L., Lauch, K., and Bird, A. (2022). ”well, actually”: investigating mansplaining in the modern workplace. *Journal of Management & Organization*, pages 1–19.
- Solnit, R. (2014). *Men explain things to me*. Haymarket Books, Chicago, IL, United States.
- Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, pages 2239–2250, New York, NY, USA. Association for Computing Machinery.

- Szlavi, A. and S. Guedes, L. (2023). Gender Inclusive Design in Technology: Case Studies and Guidelines. In *Design, User Experience, and Usability: 12th International Conference, DUXU 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23–28, 2023, Proceedings, Part I*, pages 343–354, Berlin, Heidelberg. Springer-Verlag.
- Szlávi, A. and Landoni, M. (2022). Human Computer Interaction - Gender in User Experience. In Stephanidis, C., Antona, M., and Ntoa, S., editors, *HCI International 2022 Posters*, pages 132–137, Cham. Springer International Publishing.
- Szymańska, M. (2022). Gendered neologisms beyond social media: the current use of mansplaining. *Research in Language*, 20(3):259–276.
- Söderlund, M. (2022). Service robots with (perceived) theory of mind: An examination of humans’ reactions. *Journal of Retailing and Consumer Services*, 67:102999.
- Tanksley, T. C. (2024). “We’re changing the system with this one”: Black students using critical race algorithmic literacies to subvert and survive AI-mediated racism in school. *English Teaching: Practice & Critique*, 23(1):36–56. Publisher: Emerald Publishing Limited.
- Tannen, D. et al. (1991). *You just don’t understand: Women and men in conversation*, volume 21. Virago London, New York.
- Thapar-Björkert, S., Samelius, L., and Sanghera, G. S. (2016). Exploring Symbolic Violence in the Everyday: Misrecognition, Condescension, Consent and Complicity. *Feminist Review*, 112(1):144–162. Publisher: SAGE Publications.
- Tolmach Lakoff, R. (1990). *Talking Power: The Politics of Language in Our Lives*. Basic Books, A Division of Harper-Collins Publishers.
- Trauth, E. M. (2006). *Encyclopedia of gender and information technology*. IGI Global.
- Vorvoreanu, M., Zhang, L., Huang, Y.-H., Hilderbrand, C., Steine-Hanson, Z., and Burnett, M. (2019). From gender biases to gender-inclusive design: An empirical investigation. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*, pages 1–14.
- Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, pages 1–15, New York, NY, USA. Association for Computing Machinery.

- Wang, Q., Saha, K., Gregori, E., Joyner, D., and Goel, A. (2021). Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–14, New York, NY, USA. Association for Computing Machinery.
- Wang, Q., Walsh, S., Si, M., Kephart, J., Weisz, J. D., and Goel, A. K. (2024). Theory of Mind in Human-AI Interaction. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, pages 1–6, New York, NY, USA. Association for Computing Machinery.
- Wang, Z. and Potts, C. (2019). TalkDown: A Corpus for Condescension Detection in Context. arXiv:1909.11272 [cs].
- Wardat, Y., Tashtoush, M. A., AlAli, R., and Jarrah, A. M. (2023). Chatgpt: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7):em2286.
- Wehrwein, E. A., Lujan, H. L., and DiCarlo, S. E. (2007). Gender differences in learning style preferences among undergraduate physiology students. *Advances in Physiology Education*, 31(2):153–157. Publisher: American Physiological Society.
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., and Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 214–229, New York, NY, USA. Association for Computing Machinery.
- Weiser, E. B. (2000). Gender differences in internet use patterns and internet application preferences: A two-sample comparison. *Cyberpsychology and behavior*, 3(2):167–178.
- West, S. M., Whittaker, M., and Crawford, K. (2019). Discriminating systems. *AI Now*.
- Wilson, C. C. and Gutiérrez, F. (1985). *Minorities and Media: Diversity and the End of Mass Communication*. SAGE Publications. Google-Books-ID: t8lhAAAAMAAJ.
- Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In Tang, J., Kan, M.-Y., Zhao, D., Li, S., and Zan, H., editors, *Natural Language Processing and Chinese Computing*, pages 563–574, Cham. Springer International Publishing.

- Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M. B., and Kang, B. (2023). Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Human-Centric Intelligent Systems*, 3(3):161–188.
- Ytsma, J. and Giles, H. (1997). Reactions to Patronizing Talk: Some Dutch Data. | Journal of Sociolinguistics | EBSCOhost. ISSN: 1360-6441 Issue: 2 Pages: 259 Volume: 1.
- Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., et al. (2021). The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*.
- Zhang, J., Chao, H., Dasegowda, G., Wang, G., Kalra, M. K., and Yan, P. (2022). Overlooked Trustworthiness of Saliency Maps. In Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 451–461, Cham. Springer Nature Switzerland.

APPENDICES

Appendix A

Additional Study Materials

A.1 Information Letter and Consent Form

Project Title: Understanding Explainable AI within Chatbots

Student Investigator: Natalie Nova, Management Science & Engineering
nnova@uwaterloo.ca

Faculty Supervisors:

Mark Hancock, Management Sciences, Management Science & Engineering
mark.hancock@uwaterloo.ca

Cayley MacArthur, Stratford School of Interaction Design and Business
cayley.macarthur@uwaterloo.ca

This letter explains what the study is about, possible risks and benefits, and your rights as a research participant. You may print/save a copy for your records. If you do not understand something in the letter, please ask one of the investigators before consenting to participate.

Project Summary:

A great deal of the work being done within the field of Explainable AI (XAI) is focused on answering the question of how to explain agents' behaviour. This goal has consequently dominated the landscape of Explainable AI (XAI) to focus on creating highly technical explanations that represent the internal systems of AI models. An emerging branch of XAI focuses on the human-centred approach, considering the person receiving the explanations and applying human-computer interaction (HCI) methodology to developing XAI.

The purpose of our study is to understand human and Explainable AI interaction and gain better insight into characteristics that are important to consider when designing XAI techniques.

Procedures:

To be eligible for this study, you must be 18 years old or over. Questions will ask you to provide us with your demographic information (age, gender/sex, ethnicity, occupation, education). You will be interacting with an Explainable AI system and after the interactions with the system, you will be given a questionnaire to answer with questions geared towards understanding your experience.

Your participation in this study is voluntary.

Your computer-based interactions, which includes your chat history with the chatbot within the study will also be captured and stored in a computer log file. You will not be personally identified in any study report. You may decline to respond to questions if you wish. You may withdraw your participation at any time without penalty by notifying the researchers.

You can withdraw your consent to participate and have your data destroyed by contacting us. We will keep our study records for a minimum of 8 years. Once the paper is published, data uploaded to the Open Science Framework will not be retracted. This includes the chat logs, survey responses and participant's demographic information for transparency in publication purposes.

Confidentiality and Data Security:

Your identity will be confidential. With your permission anonymous quotations may be used. In these cases participants will be referred to as Participant 1, Participant 2, ... (or P1, P2, ...) . Participant's demographic information will be shared to the Open Science Framework to encourage transparent research (no forging or manipulation of data is involved). All sensitive information will be anonymized.

You will be completing the study by an online survey operated by Qualtrics. Qualtrics has implemented technical, administrative, and physical safeguards to protect the information provided via the Services from loss, misuse, and unauthorized access, disclosure, alteration, or destruction. However, no Internet transmission is ever fully secure or error free. Qualtrics temporarily collects your computer IP address to avoid duplicate responses in the dataset but will not collect information that could identify you personally.

Prolific collects your User ID that University of Waterloo researchers use to avoid duplicate responses in the dataset, but we will not collect information that could identify

you personally. Prolific is situated in the United Kingdom and as such, your data will be temporarily stored on UK servers and the subject to EU privacy rules under the EU General Data Protection Regulation which serves to safeguard your privacy. If you prefer not to submit your online survey responses through an online platform, please do not participate in this study.

Remuneration:

Participation in this study is completely voluntary and will take approximately 10 minutes of your time. £1.50 GBP will be given to you for your time invested in this important step of this research. You may decline to answer any questions that you do not wish to answer, and you can withdraw your participation at any time by ceasing to answer questions, without penalty or loss of remuneration. To receive remuneration please proceed to the end of the questionnaire, after which you will be automatically redirected back to Prolific and receive payment. We will conduct 1 comprehension check that should you not pass, we will give you a second chance to re-try the study. This is in line with the Prolific guidelines. There will also be 2 attention checks that will determine if you have maintained attention throughout the study.

Risks and Benefits:

There are no known or anticipated risks from participation in this study. There are no direct benefits to you from participation. However, the results of this research may contribute the development of the Explainable AI space.

Research Ethics Clearance:

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB #46046). If you have questions for the Board, contact the Office of Research Ethics, toll-free at 1-833-643-2379 (Canada and USA), 1-519-888-4440, or reb@uwaterloo.ca.

For all other questions regarding this project please contact the researchers listed above. Thank you for your interest in this project.

- I agree to the use of anonymous quotations in any thesis or publication that comes of this research.
- I agree that my data can be shared in an online repository (such as Open Science Framework) as described in the information letter

By signing this consent form, you are NOT waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities.

I have read the Information Letter regarding the study being conducted by the Master's student Natalie Nova at the University of Waterloo.

I have had the opportunity to ask questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted. I am aware that explanations I provide for answers may be included in the thesis and/or publications to come from this research, with the understanding that the quotations will be anonymous. My anonymized data might be uploaded to public databases (e.g., Open Science Framework (OSF)) for future usage.

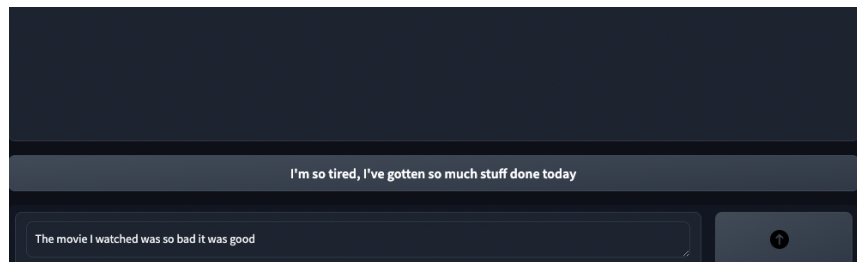
I was informed that I may withdraw my consent at any time during the study without penalty by advising the researcher.

- I am 18 or over and I agree to participate in this study
- I do not agree to participate in this study (please close your browser now)

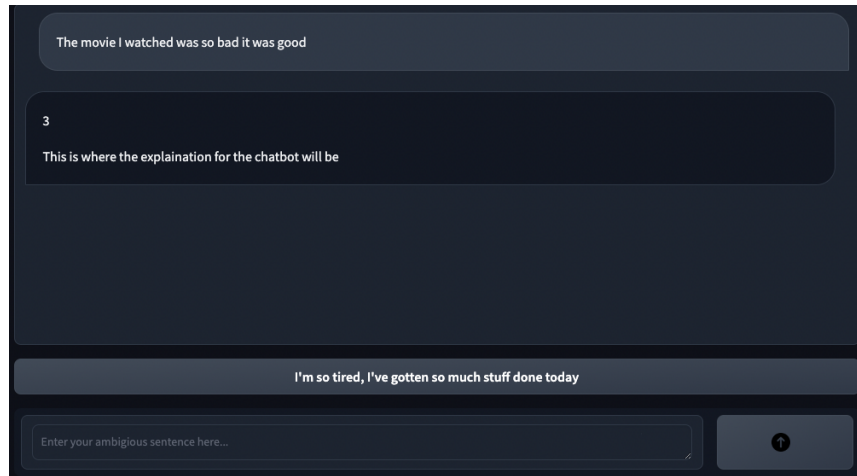
A.2 Task Instruction

Instructions: Your task is to have a conversation with the given chatbot where you will ask 3 unique ambiguous sentences. This chatbot is programmed to understand ambiguous sentences and will not understand additional instructions.

For instance, you can enter an ambiguous sentence like "The movie I watched was so bad it was good"



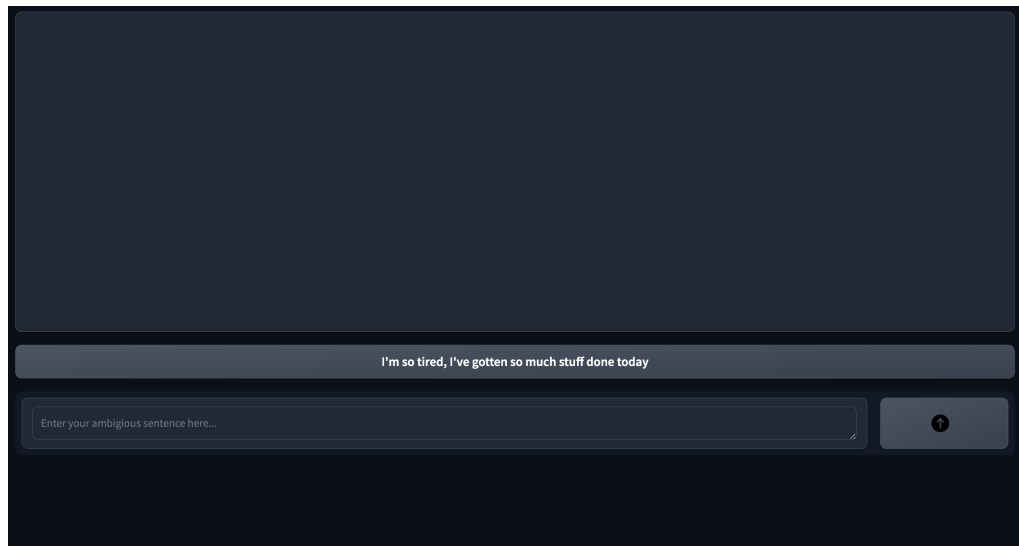
Either click enter or press the submit button and you will get an answer from the chatbot. It will evaluate and provide a sentiment score for your sentence ranging from 1 to 10, where 1 signifies a negative sentiment and 10 represents a positive one. Following the score, they will explain to justify their assessment.



When you are ready please type in your first message and wait for the chatbot's output to be displayed in full before sending a new message.

It is important to read each answer from the chatbot carefully before moving on to the next input.

Have a conversation with this chatbot where you ask **THREE** unique ambiguous sentences. Read each answer from the chatbot carefully.



Have a conversation with this chatbot where you ask **THREE** unique ambiguous sentences. Read each answer from the chatbot carefully.

Be sure that you have submitted THREE ambiguous sentences to the chatbot before moving on to the questions.

A.3 Feedback Letter

DEBRIEFING LETTER FOR STUDIES INVOLVING PARTIAL DISCLOSURE

Study Title: Manspl(AI)ning: A Gender Inclusive Approach to XAI

Student Investigator: Natalie Nova nnova@uwaterloo.ca

Faculty Supervisors: Mark Hancock, Management Sciences, mark.hancock@uwaterloo.ca
Cayley MacArthur, cayley.macarthur@uwaterloo.ca

We greatly appreciate your participation in our study and thank you for spending the time helping us with our research. When you began the study, you were told that the purpose of this study was to examine human interaction with various explainable AI techniques. However, the study was more complicated than we explained at the beginning. We wanted to conduct an exploration in human interaction with Explainable AI (XAI) but also discover what characteristics are important to consider. Historically, XAI developers make assumptions for what techniques work best for their understanding, but these assumptions can make for explanations that are not conducive to the general population. Our objective was to consider gender in the equation and to understand if there is a difference of experience with XAI techniques based on gender. We are investigating to see the effect of language and the assumption of expertise that XAI systems project on their human counterparts. We wanted to see if an experience with XAI systems can cause a similar effect to the neologism “Mansplaining”.

We could not give participants complete information about the study before their involvement because it may have influenced participants’ behaviour during the study in a way that would make investigations of the research question invalid. The reason that we used partial disclosure in this study was because we needed participants’ behaviour and attitudes to be unaffected by the study objectives. We apologize for omitting details and for providing you with fictional information about the purpose of and tasks in our study. We hope that you understand the need for partial disclosure now that the purpose of the study has been more fully explained to you. We would also like to assure you that most Human Computer Interaction research does not involve the use of partial disclosure.

We would just like to re-iterate a few things: 1. The purpose of this study was to understand whether gender is a characteristic that affects experience with Explainable AI

systems. 2. If a human's interaction with and XAI system can illicit the feeling of being mansplained to.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB #46046). If you have questions for the Board, contact the Office of Research Ethics, toll-free at 1-833-643-2379 (Canada and USA), 1-519-888-4440, or reb@uwaterloo.ca.

For all other questions contact Natalie Nova @ nnova@uwaterloo.ca If any of the questions or exercises in this study caused you to feel uncomfortable, please feel free to contact Natalie Nova, anytime by email at nnova@uwaterloo.ca. You can also contact my faculty supervisor, Mark Hancock at mark.hancock@uwaterloo.ca or Cayley MacArthur at cayley.macarthurt@uwaterloo.ca

Your identity is confidential. The data will be stored with all identifying or potentially identifying information removed. Data will be stored for at least 8 years.

Because the study involves some aspects that you were not told about before starting, it is very important that you not discuss your experiences with any other participants who potentially could be in this study. If people come into the study knowing about our specific predictions, as you can imagine, it could influence their results, and the data we collect would not be usable.

We really appreciate your participation, and hope that this has been an interesting experience for you.

A.4 Questionnaire

This section outlines the survey the participants were given once they had interacted with their chatbot.

The chatbot’s feelings towards me were:

	Strongly Disagree	Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Agree	Strongly Agree
Dominant							
Respectful							
Patronizing							
Considerate							
Warm							
Supportive							
Frustrated							
Helpless							

Table A.1: Likert scale presented for: “The chatbot’s feelings towards me were”

My feelings towards the chatbot were:

	Strongly Disagree	Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Agree	Strongly Agree
Confident							
Friendly							
Trustworthy							
Helpful							
Knowledgeable							

Table A.2: Likert scale presented for: “My feelings towards the chatbot”

I believe that the chatbot is:

	Strongly Disagree	Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Agree	Strongly Agree
Intelligent							
Incompetent							
Confident							
Unfriendly							
Helpful							

Table A.3: Likert scale presented for: “I believe that the chatbot is.”

The chatbot believes:

	Strongly Disagree	Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Agree	Strongly Agree
I am not competent							
It has greater expertise							
It is more knowledgeable							
I lack experience							

Table A.4: Likert scale presented for: “The chatbot believes.”

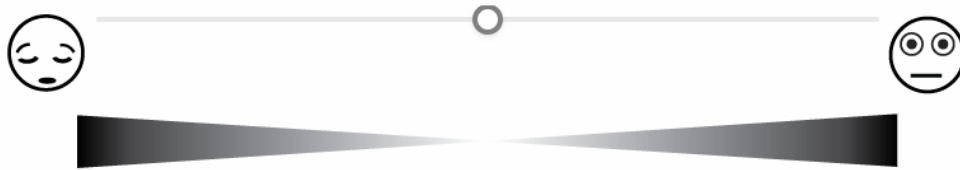
The chatbot’s language was:

	Strongly Disagree	Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Agree	Strongly Agree
Polite							
Appropriate							

Table A.5: Likert scale presented for: “The chatbot’s language was.”

Please rate your experience with the chatbot using BOTH sliders below. Don't think too much about it, just rate how you felt within the interaction.

Move your slider to rate your level of arousal



Move your slider to rate your level of pleasure

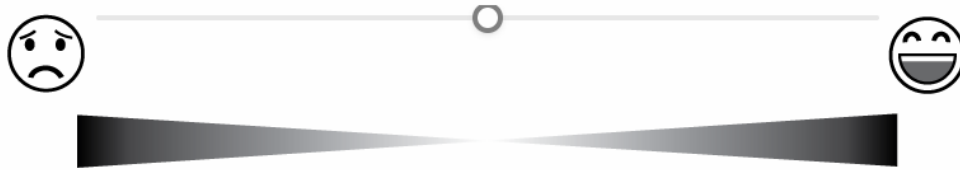


Figure A.1: The Affective Slider scale question as seen in the questionnaire

A.4.1 Overall Experience

In addition to the individual attribute ratings, participants were asked to summarize their overall experience with the chatbot by responding to the Affective Slide scale and Net Promoter Score:

How likely are you to recommend this chatbot to a friend or colleague?

Not at all likely Extremely likely

0 1 2 3 4 5 6 7 8 9 10

Figure A.2: The Net Promoter Score question as seen in the questionnaire

Additionally, participants were given an open-ended question to elaborate on their experience:

“Feel free to elaborate on any part of your interaction with the chatbot.”

This was followed by a text box for free-form responses, allowing participants to provide qualitative feedback if they chose to do so.

A.4.2 Demographics

The survey concluded with demographic questions designed to capture essential participant characteristics, which were used for contextualizing the study’s findings.

Gender identity was the main demographic consideration, with options for participants to identify as woman, man, non-binary/third gender (with an open text field for self-description), or prefer not to say. This inclusion aims to explore the influence of gender dynamics on user perceptions. Participants were further asked if they consider themselves transgender, providing options of yes, no, or prefer not to say. This question is important to understand the unique experiences of transgender individuals in human-computer interactions, which may differ from those of cisgender users.

Participants were asked to provide their age, the highest level of education completed was collected and also included a question about participants’ occupations to comprehend their professional backgrounds better. Furthermore, participants were asked to identify their race or racial background. Recognizing that racial categories are socially constructed but still impactful to facilitate intersectionality.

Lastly, participants were inquired about their frequency of use of generative AI tools, which serves to gauge their familiarity with AI technologies. This information is critical, as prior experience with AI can shape expectations and satisfaction levels in interactions.

A.5 Code

The code for this study can be found here: <https://github.com/waterloo-touchlab/Chatbots>

```
import os
import openai
import gradio as gr
from datetime import datetime
import csv
import time
import json
import jsonpickle

openai.api_key = "GOES HERE"

# This function works with js to get the window and the parameters
get_window_url_params = """
    function(text_input, url_params) {
        console.log(text_input, url_params);
        const params = new URLSearchParams(window.location.search);
        url_params = Object.fromEntries(params);
        return [text_input, url_params];
    }
    """
conversations = {}

#this function allows us to get the query strings from the url when the
study is running, this way participants don't have to add their id and it
is done manually
#INPUT: text is the variable of the query string and url_param is the value
{text = id} {url_param = 1234}
#OUTPUT: return the values that get loaded from a new page
def get_user_session_param(text, url_params):
    global raw_id
    raw_id= url_params
```

```

# parsed_json = json.dumps(raw_id)
# #print(parsed_json)
# json_string = jsonpickle.encode(parsed_json)
# recreated_obj = jsonpickle.decode(json_string)
user_id = raw_id['id']
#user_id = raw_id['id']
print(user_id)
conversations[user_id] =[
    {"role": "system",
     "content": f"PROMPT GOES IN HERE"
    }
]
return text, raw_id

```

```

#-----
#INPUT: takes in the history_open_ai formatted messages
#OUTPUT: receives chunks of openAI based on the history of the conversation
and returns those chunks as a list
def chat_completion(messages: list) -> list[str]:

    response = openai.ChatCompletion.create(
        model='gpt-3.5-turbo',
        messages=messages,
        user = '$user_id',
        stream=True
    )

    collected_messages = []
    for chunk in response:
        delta = chunk['choices'][0]['delta']
        if 'content' in delta.keys():
            collected_messages.append(delta['content'])
    return collected_messages

```

```

#INPUT: current chatbot history
#OUTPUT: chatbot history formatted into openAPI formate to feed into the

```

```

LLM based on openAI documentation
def format_messages(chat_history: list[list]) -> list[dict]:
    #global header_counter
    #global raw_id
    global user_id
    global history_openai_format
    #text, raw_id = get_user_session_param(chat_history, url_params)
    global csv_file_name
    #print("raw", raw_id)

    #user_id = parsed_json['id']

    # Check if it's a new conversation (first user message)
    #print(chat_history)
    #user_id = raw_id['id']
    if len(chat_history) == 1:
        user_id = raw_id['id']
        #user_id = raw_id['id']
        print(user_id)
        # Create a CSV file with the variable value as the name
        csv_file_name = f"{user_id}.csv"

    #sets the starting point to feed into openAI
    #prompt goes here
    history_openai_format = [
        {"role": "system",
         "content": f"PROMP HERE"
        }
    ]
    for i in range(len(chat_history)):
        ch = chat_history[i]
        history_openai_format.append(
            {
                "role": "user",
                "content": ch[0]
            }
        )

```

```

if ch[1] != None:
    history_openai_format.append(
        {
            "role": "assistant",
            "content": ch[1]
        }
    )

#Check if user does not have an existing conversation history
if user_id not in conversations:
    conversations[user_id] = []
    conversations[user_id] = history_openai_format

for key, value in conversations.items():
    if all(item in history_openai_format for item in value):
        conversations[key] = history_openai_format
        csv_file_name = f"{key}.csv"
return history_openai_format, csv_file_name

```

```

# This function deals with taking all the information combining it and
writing it into a csv file
#INPUT: history, csv file name which we want to write into and the most
recent chatbot answer
#OUTPUT: no return value but the output is an appended csv file with the
most recent conversation in the format of history_openai_format aka
[dictionaries] def write_file(history_openai_format, csv_name,
chatbot_answer):
    # Append the chat history to the CSV file with the header
    if chatbot_answer[0][1] != None:
        final_output = history_openai_format

        final_output.append(
            {
                "role": "assistant",
                "content": chatbot_answer[-1][1]
            }
        )

```

```

)
with open('/filePath/'+ csv_name, mode='a', newline='') as f:
    writer = csv.writer(f)
    # Get the current timestamp
    current_time = datetime.now().strftime("%Y-%m-%d %H:%M:%S")
    # Write a new header for each new conversation
    writer.writerow([f"Header {current_time}"])
    # Write the chat history for this conversation with the
    conversation ID
    writer.writerow(final_output)

#INPUT: recieves chatbot(history) from set_user_query
#OUTPUT: the text for the chatbot answer
def generate_response(text: str, chatbot_answer: list[list]):
    formatted_history, csv_name = format_messages(chatbot_answer)
    #this value is going to hold a list of chunks coming in from openAI
    as a text response
    bot_messages = chat_completion(formatted_history)

    #last answer in conversation
    chatbot_answer[-1][1] = ''
    #streams the output so it looks like chatgpt
    #for each chunk add to the answer to stream the output
    for bm in bot_messages:
        chatbot_answer[-1][1] += bm
        time.sleep(0.05)
        #yield is a function in gradio that lets your stream the answer
        yield chatbot_answer

    write_file(formatted_history, csv_name, chatbot_answer)
    return chatbot_answer

# sets the input from the user before the output gets processed
#INPUT: text of a string from the textbox which is the user message
and chatbot which is the history (a list of lists)
#OUTPUT: Sets user response and waits for chatbots answer

```

```

def set_user_query(text: str, chatbot: list[list]) -> tuple:
    #append the text of the user to the chatbot history and set chatbot
    response to none
    chatbot += [[text, None]]
    return '', chatbot

# GRADIO UI
with gr.Blocks() as demo:
    chatbot = gr.components.Chatbot(label='Assistant', show_label = False)
    with gr.Row(equal_height=True, variant= 'panel'):
        msg = gr.components.Textbox(scale = 6, placeholder="Enter your
            ambiguous sentence here...", show_label = False)
        button = gr.components.Button(value = " ", variant = 'secondary',
            scale = 0, icon = "send_icon.png")
    url_params = gr.JSON({}, visible=False, label="URL Params")

    print(url_params)

    msg.submit(
        fn=set_user_query,
        inputs=[msg, chatbot],
        outputs=[msg, chatbot],
    ).then(
        fn=generate_response,
        inputs=[msg, chatbot],
        outputs=[chatbot]
    )

    button.click(
        fn=set_user_query,
        inputs=[msg, chatbot],
        outputs=[msg, chatbot]
    ).then(
        fn=generate_response,
        inputs=[msg, chatbot],
        outputs=[chatbot]
    )

```

```
demo.load(  
    fn=get_user_session_param,  
    inputs=[msg, url_params],  
    outputs=[chatbot, url_params],  
    _js=get_window_url_params  
)  
  
demo.queue()  
demo.launch()
```