F21CA Coursework Group Report

Conversational Agents and Spoken Language Processing

Alignment with Social Robots

Supervisors:

Alessandro Suglia
Marta Romeo
Greta Gandolfi

Group members:

Aswin Shaji	- **as2204**
Guilhem Santé	- **gs2061**
Jeff Sherer	- **jjls2000**
Jake Paterson	- **jp116**
Maximilien Bildstein	- **mb2156**
Mizan Haque	- **mmsh2000**
Marc Puig Arocas	- **mp2090**
Prudhvi Venkata Sai Paleti	- **pp2054**
Robbie McPherson	- **rm2130**
Sattwik Mohanty	- **sm2163**

April 2024

Abstract

This study investigates the impact of integrating anthropomorphic traits and personality-prompted Large Language Models (LLMs) into robots on social alignment in human-robot interactions (HRI). This paper conducted a two-phase experiment with participants interacting with a Furhat robot, revealing significant increases in alignment correlation scores post-interaction across both control and prompt conditions. Notably, improvements in social alignment were consistent regardless of the nature of the task. The findings suggest that engineered personality traits play a critical role in shaping human perceptions and enhancing engagement in HRI, with implications for sectors such as customer service, healthcare, and education. Despite constraints in sample size and participant diversity, this research lays a foundation for future exploration of personality customisation and its impact on social alignment dynamics, paving the way for more intuitive and effective human-robot relationships in diverse application domains. Ethical considerations were prioritised throughout data collection and analysis to ensure participant confidentiality and anonymity.

1 Introduction

In interpersonal communication, social alignment is crucial for shaping meaningful interactions. Garrod and Anderson (1987) emphasise effective communication, stressing the need to align language use and interpretation beyond vocabulary alone, considering factors like sentence structure, meaning, and contextual cues (Pickering and Garrod, 2004, 2021). Achieving this alignment is essential for fostering productive dialogue, mutual understanding, and cooperation. Moreover, social alignment extends beyond human-to-human interactions, particularly in the realm of social robotics and human-robot interactions (HRI). The recent surge in Large Language Models (LLMs) and the development of advanced anthropomorphic humanoid social robots further underscore the importance of social alignment in HRI (Duffy, 2003). However, this convergence of social psychology, LLMs, and social robots is still in its infancy.

Exploring the intricate nature of social alignment, as highlighted by Gallotti et al. (2017), unveils its dynamic and interactive qualities, which intricately shape the exchange of mental attitudes and representations during social interactions. This complexity in social alignment poses notable challenges in measurement, particularly within the domain of HRI (Mayima et al., 2022).

1.1 Aims and Objectives

The aim of this paper is to anthropomorphise a Furhat robot with a custom personality trait, prompt-engineered LLM, and examine its influence on social alignment. To achieve this, the paper's objective is to measures social alignment by assessing individuals' beliefs before and after interacting with the robot, in order to evaluate any shifts or changes in their mental attitudes and representations. Through the development and implementation of this methodology, the goal of this paper is to offer insights into the dynamics of social alignment in HRI, thereby contributing to advancements in HRI research.

Figure 1 demonstrates various Furhat robots, each featuring distinct facial anthropomorphic qualities.

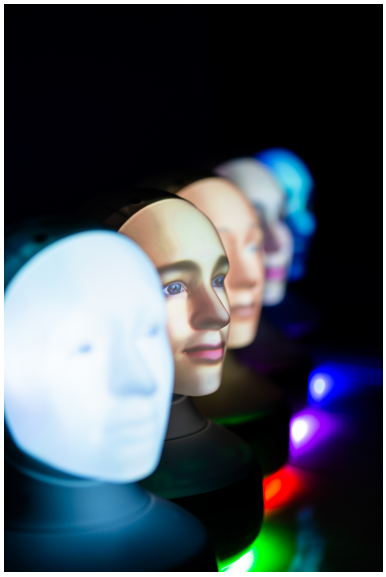

Figure 1: A collection of Furhat Robots with different facial anthropomorphic qualities.

Source: https://furhatrobotics.com/furhat-robot/

2 Literature Review

2.1 Background

In her book, Breazeal (2004) envisioned a seamless interaction between humans and robots, emphasising the importance of making robots more lifelike to improve social interaction (Kühne and Peter, 2023). This aligns with the concept of anthropomorphism in HRI, where human traits are

projected onto robots (Kühne and Peter, 2023). Inspired by human communication, technology design has focused on enhancing robot usability in social settings (Guzman, 2018), resulting in efforts to make robots more human-like in their communication and behaviour (Breazeal, 2003). These advancements are aimed at fostering meaningful social interactions and relationships (Fong et al., 2003).

In HRI, social alignment plays an important role in the development of robots capable of effective communication and meaningful interaction with humans, prompting researchers to investigate the factors influencing alignment tendencies. In their work, Branigan et al. (2011) discovered that participants exhibited a stronger inclination to align with 'computer' partners compared to 'human' partners, especially when the computers were portrayed with varying levels of capabilities. This highlights how alignment tendencies in human-robot interaction can be influenced by individuals' beliefs and the communicative abilities of their interaction partners.

A strong measure of communicative proficiency involves the capacity to convey complex ideas effectively, demonstrate active listening skills to comprehend differing viewpoints, and foster dialogues that promote transparent information exchange. In social psychology, effective communication and the sharing of knowledge are integral components of interpersonal interaction and group dynamics (Kozlowski and Ilgen, 2006). Studies rooted in personality psychology have indicated that individuals' levels of openness, conscientiousness, agreeableness, and emotional stability, as defined by the Big Five personality traits, can significantly influence their communication styles and ability to share knowledge (Soto and Jackson, 2013).

The integration of LLMs in HRI settings has led to novel opportunities in prompt engineering, enabling researchers to design prompts that evoke specific responses and shape human-robot interactions, particularly in exploring the effects of anthropomorphism on user perceptions and behaviours. In their article, Jiang et al. (2024) developed 'Personality Prompting' as a way to inject anthropomorphic personality traits into LLMs in a 'controllable' way. This ability to prompt personalities raises the following question: Is it possible to prompt engineer specific personality traits into LLMs and integrate them with social robots to facilitate knowledge sharing and foster social alignment in human-robot interaction?

After thorough research, the authors found no existing papers or findings that tackle this question. Therefore, they consider this paper to be the first to address it.

2.2 Psychological Dynamics and Personality Traits in Social Interaction

Social proof and cognitive dissonance are fundamental elements within social interactions, influencing individuals' perceptions and behaviours in various contexts (Vashistha et al., 2018; Vargheese et al., 2020; Ibrahim et al., 2013; Matz and Wood, 2005; Sukmayadi and Yahya, 2020). Initially conceptualised by Festinger (1957), cognitive dissonance describes the discomfort stemming from conflicting beliefs, prompting individuals to seek consistency by adjusting their beliefs or rejecting new information. Complementing this, social proof, as outlined by Cialdini (1984), operates through conformity, as individuals align their actions and responses with those around them. Within social contexts, social proof validates new information or behaviours, facilitating their acceptance. Understanding the interplay between cognitive dissonance and social proof aids communicators in effectively framing information and leveraging social dynamics to foster knowledge exchange.

This psychological backdrop intersects with research on personality traits, particularly within the Big Five framework explored by Yin et al. (2023), which examines their impact on knowledge-sharing capabilities. Knowledge sharing hinges on individuals' confidence in contributing relevant knowledge and skills (Ozer and Reise, 1994), while the Big Five personality traits—agreeableness, emotional stability, conscientiousness, openness to experience, and extraversion—influence how individuals engage in such interactions (Ozer and Reise, 1994). Extraversion stands out as the primary trait influencing knowledge sharing, with extraverted individuals, known for their sociable and outgoing demeanour, demonstrating proficiency in initiating and seizing opportunities for knowledge exchange, as well as effectively discerning the needs and knowledge of others (Yin et al., 2023).

As aforementioned, Jiang et al. (2024) introduces a pioneering methodology, Personality Prompting (P^2), aimed at inducing specific personalities in LLMs. Grounded in psychological re-

search, P^2 employs a sequential prompt-generation process to control LLM behaviours, aligning them with desired personality traits. It does so by employing an iterative chain-of-thought prompting method that facilitates the development of the desired personality traits through a series of intermediate reasoning steps (Wei et al., 2022). Additionally, Jiang et al. (2024) introduces the concept of a "question prompt," a tool that contextualises the prompted personality trait under investigation. In line with the project's aims and objectives, the P^2 method enables engineers to explore and customise how personality traits, in conjunction with contextual factors such as cognitive dissonance and social proof, influence knowledge exchange and social alignment. By understanding how individuals reconcile conflicting beliefs and conform to social norms, engineers can effectively leverage the P^2 method to investigate the intricate dynamics shaping communication and alignment in HRI.

2.3 Integration of LLM Technologies in Emerging HRI Domains

LLMs are poised to have a significant impact across various sectors, particularly in roles reliant on information processing and generation, as expressed in OpenAI's report on their potential labour market effects (Eloundou et al., 2023). The indication of social alignment presents opportunities to enhance human-robot interaction, exemplified by integrating anthropomorphic traits and prompt-engineered LLMs with social robots (Zhang et al., 2023). Embodied robots, equipped with extensive communication capabilities, encourage higher levels of engagement and trust compared to text-based or virtual agents (Kim et al., 2024). LLMs enable robots to understand natural language instructions and engage in adaptable dialogue interactions, facilitating tasks like planning and collaboration (Kim et al., 2024). Recent advancements, including PaLM-E and LM-Nav, demonstrate LLMs' proficiency in bridging language and perception gaps, enabling autonomous instruction generation and effective user communication (Zeng et al., 2023). This convergence of LLMs and embodied robots holds promise for improving social alignment and collaborative interactions in various application domains. However, it's crucial to manage the integration of LLMs carefully, balancing technological advancements with the development of human capabilities and oversight.

3 Methodology

3.1 Participants

Postgraduate students from Heriot-Watt University's F21CA (Conversational Agents and Spoken Language Processing) program were recruited for the experiment. Recruitment was done by first establishing contact with group leaders and then agreeing on mutual participation in each group's experiments. Additionally, participants from outside the F21CA course were also recruited. No restriction were imposed on the demographics of the participants that were recruited, but an inquiry was made at the beginning of the experiment regarding participant background via a questionnaire. This was done to gain deeper insights into potential mediating variables. The data collected was participants' fluency in English, whether they had interacted with a Furhat robot in the past or not, and if they had any experience in the professions used in one of the experimental tasks to identify potential biases. Among the participants, only four had previous experience with the Furhat robot, while seven out of the participants had backgrounds in teaching, which was among one of the asked occupations. Five participants were native English speakers. (note: The collection of data from four participants was not confirmed; therefore these numbers on the participant background are only representative of 22 participants instead of 27).

3.2 Experimental design

A 2 x 2 factorial, between-subjects experiment was conducted to test the hypotheses of the study. It consisted of two phases: before an interaction with a robot, and after the interaction.

The participants were randomly assigned to two different conditions: a baseline condition and an experimental (Prompted) condition. Both conditions include a robot integrated with an LLM, modified for enhanced context management. There was a baseline condition with no additional modifications (Condition 1), and an experimental condition with a personality prompt to guide the conversational responses of the LMM (Condition 2).

Participants were randomly assigned to one of two conditions, with each participant engaging in discussions on two categories of concepts during their interaction with the robot. These categories formed the two levels of the experiment: factual and contestable. The factual concepts required ordering items based on objective, verifiable data

such as the proximity of planets from the Sun (Mercury, Venus, Earth, Mars, and Jupiter) and geometric shapes by the number of sides from least to most (triangle, square, pentagon, hexagon, and heptagon). The contestable concepts involved rankings based on subjective evaluations, including rankings on the most important careers for the future from least to most important (pilot, lawyer, farmer, artist, and teacher) and ranking of factors that contribute to a happy life from least to most important (family and friends, health, freedom, religion, and money).

The independent variable for the study was the participant assignment of a baseline or prompted robot for the interaction. The dependent variable was the change in participants' ranking correlation with the robot, from before and after the interaction.

3.3 Procedure

The experimental trials were conducted in a private room inside the National Robotatrium located on Heriot Watt University's Edinburgh campus. Consistent environmental settings were maintained to mitigate any factors that could influence the experiment. This included the same room for each experiment, the use of blinds to control lighting and potential interruptions, limited interaction with the researchers during the experiment, and a scripted introduction to the experiment was read by the researchers.

Upon arrival, participants were delivered a scripted presentation from the researchers that gave an overview of the study's objectives. Participants were then given a consent form to sign which assured adherence to ethical practices. Participants were then given another form on background information relevant to the experiment.

Next, participants were asked to complete four ranking-tasks covering the factual and contestable concepts. Following this, participants were then read another scripted message that described an overview of the next phase of the experiment including its structure and guidelines of the robot interaction.

During the structured interaction phase, participants engaged with the robot, discussing each ranking-task concept within designated time frames (2 minutes and 30 seconds per task; 10 minutes in total). Following this interaction, participants revisited the ranking-tasks and completed them again.

Following the interaction phase, participants were invited to complete the Godspeed survey on their attitudes toward the robot.

3.4 Hardware

The experiments employed a robot from Furhat Robotics known for its wide range of social features, including facial expressions, proximity sensors, and anthropomorphic traits. These features collectively improve its suitability as a model for measuring social alignment. The Furhat robot also has native features that automatically transcribe the audio of the interaction through the Google Speech-to-Text API, which is controlled remotely through HTTP requests by a running client, via the client's API.

3.5 Software Architecture

The research team developed their own client for the robot using Kotlin and built it with Gradle. The choice of Kotlin as the main programming language for this project was motivated by its advanced library dedicated to Furhat behavioral programming.

The client was designed to fulfill various project needs, including requesting response generation from the API of a popular LLM, managing the conversational state of the robot, and monitoring participant interactions with the system.

The response generation was completed through the OpenAI API using the gpt-3-turbo-instruct model. This model was - at the time of the project implementation - one of the most easily accessible instruction-tuned models, offering the shortest inference time. And while instruction-tuning reduced the context window of the LLM, it also meant the model was more capable of following the prompt engineering techniques employed. Figure 2 illustrates the architecture of the developed robot system.

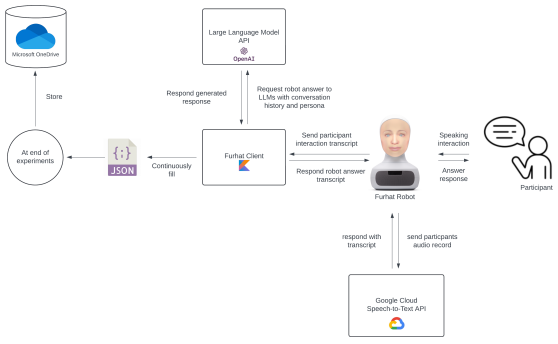

Figure 2: Overview of System Architecture

3.6 Prompt Engineering

The study utilises the Personality Prompting (P^2) (Jiang et al., 2024) method to enhance the LLM by reflecting human-like personalities in the responses of a conversational agent. Initially, a naive prompt is created to illustrate typical behavior of a desired Big Five trait (Norman, 1963). This is then refined into a keyword prompt enriched with descriptive words from psychological research to increase the saliency of the chosen Big Five trait and relevant psychology research that explores promoting social alignment. Finally, a "chain-of-thought" technique is utilised, inspired by Wei et al. (2022), to provide the conversational agent an artificial rationale for its responses. This method leverages established relationships between language use and personality of the Big Five trait chosen (Mehl et al., 2006).

3.7 GodSpeed Questionaire Series

The Godspeed Questionnaire Series (GQS) is crucial in HRI research, offering a standardised measurement of key concepts such as anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety (Bartneck, 2023). In this paper, the GQS was used to assess participants' perceptions of the anthropomorphised Furhat robot's social attributes. The questionnaire provided a structured method for gathering qualitative data on participants' impressions of the robot's personality, adding an extra layer of insight for analysis. It enabled researchers to explore participants' impressions and insights more deeply, particularly regarding their comfort levels and trust in the robot (Bartneck, 2023). These aspects are crucial for fostering effective human-robot interactions and alignment (Kelch et al., 2024). Incorporating the GQS allowed for comparisons between subjective evaluations and objective measures of social alignment, validating the effectiveness of the experimental setup and the prompt engineering strategies (Bartneck, 2023).

3.8 Data Analysis

The study will perform a statistical analysis on the change in participant ranking correlations with that of a robot that utilizes customised rankings and opinions. The investigation will also analyse how participant background factors may moderate these effects. Initially, the analysis will include a linear mixed model that will incorporate aspects such as time and conditions, and random effects for participant-specific variations. This approach helps to accurately account for the variability both within and between participants. The exploratory analysis will segment Godspeed survey results by demographic subgroups to examine the differences in perceptions of the robot. All data pre-processing and analyses were performed using the R statistical software.

4 Results

4.1 Primary Statistical Analysis

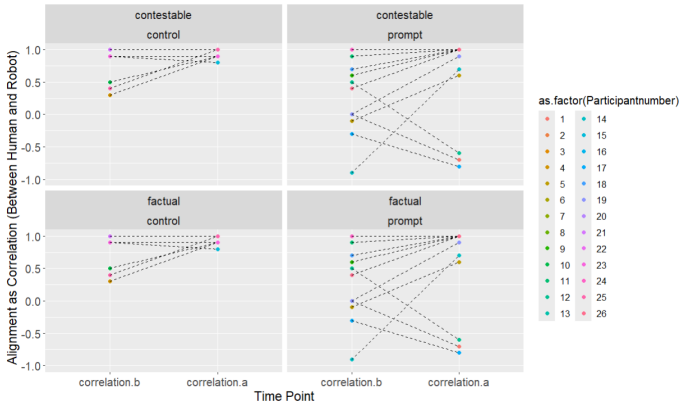

Figure 3: Correlational scores split by concept category and condition

Figure 3 illustrates the differences and similarities in the scores across various conditions. The statistical analysis demonstrated increases in post-interaction correlation scores across all conditions, with mean scores in the control condition rising from 0.82 (SD=0.25) to 0.95 (SD=0.07) and in the prompt condition from 0.41 (SD=0.57) to 0.58 (SD=0.69). The analysis on concept category for the ranking task revealed that both contestable and factual tasks had results that were highly alike, regardless of condition.

The linear mixed-effects model revealed a significant temporal effect on correlational scores, with an estimated increase of +3.984 (SE = 0.522, degrees of freedom = 178, t-value = 7.631, $p < .001$), irrespective of the condition or task type. A significant interaction between condition and time (Estimate = +4.692, SE = 1.044, df = 178, $p < .001$) was observed, indicating that the increases in alignment were more pronounced when under different conditions. Despite this significant interaction between condition and time, the condition effect on its own was not significant (Estimate = -3.120, SE = 3.138, df = 24, p = .33). Figure 4 neatly summarises the statistical influences observed in the

analysis, including the significant temporal effect and the interaction between condition and time on correlational scores.

The REML criterion value of 1216 for our linear mixed-effects model suggests a good fit, with residuals normally distributed around the median (-0.0707). Variability across participants was significant as shown by a high variance in intercepts (Variance = 61.87, SD = 7.866).

	Alignment: Z-transformed Correlation Scores		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	8.92	5.82 – 12.01	**<0.001**
condition1	-3.12	-9.31 – 3.07	0.321
Time Point [After Discussion]	3.98	2.95 – 5.01	**<0.001**
type1	-0.00	-1.03 – 1.03	1.000
condition1:time1	4.69	2.63 – 6.75	**<0.001**
time1:type1	0.00	-2.05 – 2.05	1.000
Random Effects			
σ^2	14.09		
$\tau_{00\ Participantnumber}$	61.87		
$\tau_{11\ Participantnumber.type1}$	0.00		
$\rho_{01\ Participantnumber}$	-1.00		
$N_{\ Participantnumber}$	26		
Observations	208		
Marginal R^2 / Conditional R^2	0.367 / NA		

Figure 4: Estimated Effects of Fixed Factors Using a Linear Mixed Model

4.2 Exploratory Analysis

The Godspeed survey showed overall positive ratings for the Furhat robot across categories such as naturalness, human-likeness, and consciousness. Subgroup analysis revealed nuanced differences: Native English speakers perceived the robot as less natural and human-like (Mean = 3.0 and 2.6) compared to non-natives, but rated high in competence (Mean = 4.4). Intermediate English speakers generally viewed the robot as engaging and competent, with ratings ranging from 3.33 to 4.56. Advanced speakers also consistently rated the robot's human-like qualities high (Means > 4.0). Participants with prior robot interactions also reported relatively high ratings across all attributes (Means from 3.75 to 5.0). Conversely, those without prior experience gave moderate to high ratings (3.22 to 4.5).

5 Discussion

This study focused on the impact of anthropomorphism of a robot with a custom prompt-engineered LLMs on social alignment in human-robot interactions. Utilising a two-phase experimental design, our findings revealed significant increases in alignment correlation scores from pre-interaction to post-interaction across both conditions, indicating the occurrence of social alignment after the participant-robot interaction.

The results indicated an increase in alignment scores post-interaction in both the control and prompt conditions, with more variability in the prompt condition. This variability suggests an element of unpredictability in participant responses, particularly for the participants who interact with an enhanced, personality prompted robot. The analysis further highlighted consistent indications of social alignment across different conceptual categories (factual and contestable), indicating that the nature of the task may not determine the extent to which social alignment can occur. Alternatively, an explanation for these results could be centered in the design of the experiment. The selection of factual and contestable concepts might not have adequately captured the intended differences between categories. These concepts were chosen based on the expected willingness of participants to modify them. However, the results suggest that the distinction between concepts may be more nuanced than initially assumed, particularly in topics such as future career and contributing factors to a happy life. The latter topic will likely include emotional connections to the answers that will be unlikely to be altered after a short interaction with the robot.

This study advances our theoretical understanding of the mechanisms through which personality prompts in LLMs can enhance social alignment in HRI. By demonstrating significant improvements in alignment scores using personality prompts, our findings lend empirical support to theoretical models advocating for the anthropomorphic design in robotics to improve engagement and interaction. The ability to customise robot interactions through personality prompts could lead to more effective and meaningful human-robot interactions across various applications. Especially those that utilise robots as a form of intervention to moderate user behavior. This customisation could help tailor interactions to individual users, or companies, preferences' and needs'.

Our analysis shows that the use of personality prompts enhances social alignment, yet the effect varies across participant demographics. The ex-

ploratory analysis of the Godspeed survey revealed that perceptions of the robot's naturalness, human-likeness, and competence differ based on participants' language proficiency and prior experience with robots. The findings of the exploratory analysis illustrate the complex interplay between robot design, participant background, and interaction context. They highlight the importance of considering diverse user groups when designing and implementing personality prompts in social robots to optimise user experience and effectiveness.

6 Limitations

The project's limitations can be categorised into two primary areas: interaction constraints and evaluation limitations. As emphasised in the introduction, social alignment is a nuanced and multifaceted concept, encompassing various measurement approaches. Consequently, to ensure the feasibility of the experiment, the researchers considered and imposed specific limitations.

The study narrowed its focus to extraversion as the sole personality trait, simplifying the analysis to explore its influence on social alignment while overlooking the complexity of individuals' personalities, typically characterised by multiple traits from the Big Five model. This decision aimed to reduce confounding factors and facilitate a more targeted investigation. Additionally, the Furhat robot used default facial expressions and vocal features without customisation to avoid biasing participants' perceptions, ensuring consistency with the experiment's objectives.

The chosen evaluation plan to understand social alignment in HRI constrained our work in several ways. Most importantly, the sample size and participant diversity may have contributed to the lack of more conclusive results, as outlined by Bethel and Murphy (2010). The chosen demographic of participants was limited and may have not been sufficient for the aim of the project. This was due to resource allocation and scope strategy for the project. In terms of metrics, limiting each ranking-task to only five items per concept might not have sufficiently captured the complex dynamics of social alignment when combined with prompted LLMs in HRI. This might have simplified the subtle differences in correlation changes between user and robot rankings. Moreover, the study's emphasis on ranking-tasks, along with the broad concept categories, could have constrained the conclusions.

7 Future Work

Future research could further explore a broader spectrum of personality traits and customisation options to better understand their impact on social alignment dynamics in HRI, improving understanding of these complex interactions. For instance, many individuals exhibit combinations of multiple Big Five personality traits.This exploration could have the potential to unveil nuanced patterns in participants' perceptions and behaviours, enhancing the depth of analysis in subsequent studies. Exploring the use of advanced natural language processing techniques to refine the customisation of robot personalities could further enrich the understanding of social alignment. Specifically, future work could investigate how updating prompts to better accommodate individuals' diverse backgrounds can enhance the specificity and effectiveness of interactions.

8 Conclusion

This study has demonstrated that anthropomorphic robots can enhance social alignment in human-robot interactions. Our findings reveal improved correlations in alignment when personality-prompted LLMs are utilised. This study supports models advocating for human-like robot customisation for enhancing engagement and suggests profound practical implications for designing social robots that can better meet diverse human needs across various sectors. Overall, our research indicates that personality customisation in robots can lead to an enhanced human-robot interaction.

9 Ethical considerations

Participant rankings and background information were collected via Microsoft Forms on computers provided by the research team. Confidentiality and anonymity was established by identifying participants solely through unique numbers, instead of names, on the Microsoft Forms questionnaire. The conversation with the participant was also monitored and collected by storing the data in JSON format using the gjson library, while the interaction was simultaneously taking place. All data was securely stored using university-approved guidelines and in accordance with the UK's General Data Protection Regulation (GDPR).

References

Christoph Bartneck. 2023. *Godspeed Questionnaire Series: Translations and Usage*, pages 1–35.

Cindy L. Bethel and Robin R. Murphy. 2010. Review of human studies methods in hri and recommendations. *International Journal of Social Robotics*, 2:347–359.

Holly P Branigan, Martin J Pickering, Jamie Pearson, Janet F McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: evidence from dialogs with humans and computers. *Cognition*, 121(1):41–57.

Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems*, 42(3-4):167–175.

Cynthia Breazeal. 2004. *Designing sociable robots*. MIT press.

Robert B. Cialdini. 1984. *Influence: The Psychology of Persuasion*. William Morrow and Company, New York, NY.

Brian Duffy. 2003. Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42:177–190.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models.

Leon Festinger. 1957. *A Theory of Cognitive Dissonance*. Stanford University Press.

Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166.

M. Gallotti, M.T. Fairhurst, and C.D. Frith. 2017. Alignment in social interactions. *Consciousness and Cognition*, 48:253–261.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

A. L. Guzman. 2018. What is human–machine communication, anyway? In A. L. Guzman, editor, *Human-machine communication: Rethinking communication, technology, and ourselves*, pages 1–28. Peter Lang.

N. Ibrahim, M.F. Shiratuddin, and K.W. Wong. 2013. Persuasion techniques for tourism website design.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

Yngve Kelch, Annette Kluge, and Laura Kunold. 2024. Would you trust a robot that distrusts you? In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 588–592, New York, NY, USA. Association for Computing Machinery.

Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. 2024. Understanding large-language model (llm)-powered human-robot interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24. ACM.

Steve WJ Kozlowski and Daniel R Ilgen. 2006. Enhancing the effectiveness of work groups and teams. *Psychological science in the public interest*, 7(3):77–124.

Rinaldo Kühne and Jochen Peter. 2023. Anthropomorphism in human–robot interactions: a multidimensional conceptualization. *Communication Theory*, 33(1):42–52.

David C. Matz and Wendy Wood. 2005. Cognitive dissonance in groups: The consequences of disagreement. *Journal of Personality and Social Psychology*, 88(1):22–37.

Amandine Mayima, Aurélie Clodic, and Rachid Alami. 2022. Towards robots able to measure in real-time the quality of interaction in hri contexts. *International Journal of Social Robotics*, 14(3):713–731.

Matthias R. Mehl et al. 2006. Are women really more talkative than men? *Science*, 317(5834):82.

Bob T. Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 66:574–583.

Daniel J. Ozer and Steven P. Reise. 1994. Personality assessment. *Annual Review of Psychology*, 45(Volume 45, 1994):357–388.

Martin Pickering and Simon Garrod. 2021. *Understanding Dialogue: Language Use and Social Interaction*.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Christopher J Soto and Joshua J Jackson. 2013. Five-factor model of personality. *Journal of Research in Personality*, 42:1285–1302.

Vidi Sukmayadi and Azizul Yahya. 2020. A review of cognitive dissonance theory and its relevance to current social issues. *MIMBAR Jurnal Sosial dan Pembangunan*, 36.

John Paul Vargheese, Matthew Collinson, and Judith Masthoff. 2020. Exploring susceptibility measures to persuasion. In *Persuasive Technology. Designing for Future Change*, pages 16–29, Cham. Springer International Publishing.

Aditya Vashistha, Fabian Okeke, Richard Anderson, and Nicola Dell. 2018. 'you can always do better!": The impact of social proof on participant response bias. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA. Association for Computing Machinery.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Kui Yin, Dongfang Li, Xiaodan Zhang, Niannian Dong, and Oliver J. Sheldon. 2023. The influence of the big five and dark triad personality constructs on knowledge sharing: A meta-analysis. *Personality and Individual Differences*, 214:112353.

Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S. Yu. 2023. Large language models for robotics: A survey.

Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. 2023. Large language models for human–robot interaction: A review. *Biomimetic Intelligence and Robotics*, 3(4):100131.

Appendices

A tibble: 16 × 5
 # Groups: concept, condition [8]

	concept	condition	time	Mean	SD
	<fct>	<fct>	<fct>	<dbl>	<dbl>
1	contestable1	control	correlation.b	0.82	0.26
2	contestable1	control	correlation.a	0.95	0.07
3	contestable1	prompt	correlation.b	0.41	0.59
4	contestable1	prompt	correlation.a	0.58	0.71
5	contestable2	control	correlation.b	0.82	0.26
6	contestable2	control	correlation.a	0.95	0.07
7	contestable2	prompt	correlation.b	0.41	0.59
8	contestable2	prompt	correlation.a	0.58	0.71
9	factual1	control	correlation.b	0.82	0.26
10	factual1	control	correlation.a	0.95	0.07
11	factual1	prompt	correlation.b	0.41	0.59
12	factual1	prompt	correlation.a	0.58	0.71
13	factual2	control	correlation.b	0.82	0.26
14	factual2	control	correlation.a	0.95	0.07
15	factual2	prompt	correlation.b	0.41	0.59
16	factual2	prompt	correlation.a	0.58	0.71

Figure 5: Mean Correlation Scores and Standard Deviations of Conditions and Levels

A tibble: 8 × 5
 # Groups: type, condition [4]

	type	condition	time	Mean	SD
	<fct>	<fct>	<fct>	<dbl>	<dbl>
1	contestable	control	correlation.b	0.82	0.26
2	contestable	control	correlation.a	0.95	0.07
3	contestable	prompt	correlation.b	0.41	0.58
4	contestable	prompt	correlation.a	0.58	0.69
5	factual	control	correlation.b	0.82	0.26
6	factual	control	correlation.a	0.95	0.07
7	factual	prompt	correlation.b	0.41	0.58
8	factual	prompt	correlation.a	0.58	0.69

Figure 6: Mean Correlation Scores and Standard Deviations of all Contestable and Factual concepts

A tibble: 4 × 4
 # Groups: condition [2]

	condition	time	Mean	SD
	<fct>	<fct>	<dbl>	<dbl>
1	control	correlation.b	0.82	0.25
2	control	correlation.a	0.95	0.07
3	prompt	correlation.b	0.41	0.57
4	prompt	correlation.a	0.58	0.69

Figure 7: Mean Correlation Scores and Standard Deviations of each experimental condition

A tibble: 4 × 4
 # Groups: type [2]

	type	time	Mean	SD
	<fct>	<fct>	<dbl>	<dbl>
1	contestable	correlation.b	0.6	0.5
2	contestable	correlation.a	0.75	0.54
3	factual	correlation.b	0.6	0.5
4	factual	correlation.a	0.75	0.54

Figure 8: Mean Correlation Scores and Standard Deviations of each concept category