
Proceedings Paper:
Jones, Charlotte, Reed, Darren James orcid.org/0000-0001-9018-0145 and Camara, Fanta (2024) Who Is The Chameleon? A party game to explore trust and biases towards Alexa, Pepper and ChatGPT. In: 25th Annual Conference Towards Autonomous Robotic Systems 2024. TAROS 2024, 21-23 Aug 2024, Brunel University. Springer , GBR , pp. 175-186.

Who Is The Chameleon? A party game to explore trust and biases towards Alexa, Pepper and ChatGPT***

Charlotte Jones[a], Darren Reed[b] and Fanta Camara[c]

[a] School of Natural Sciences, University of York, UK
[b] Department of Sociology, University of York, UK
[c] Institute for Safe Autonomy, University of York, UK

Abstract. With the increasing adoption of robotic and AI systems in our daily activities, understanding how people interact with and trust these systems or may have biases towards them is becoming important for their development and safety. This paper investigates people's level of trust and potential biases towards three robotic systems i.e. an Alexa device, robot Pepper and ChatGPT using "The Chameleon" game. In this experiment, the same words were presented to two of groups of participants, the first group played the game through an online form and the second group played the game in person with Alexa, Pepper and ChatGPT in a lab space. The game consisted in spotting the player who is the chameleon i.e. pretending to know a target word that only the other players are supposed to know. The results showed that both the online and in-person participants had similar levels of spotting the chameleon. However, participants were less able to spot Pepper when it was the chameleon, suggesting that they trusted Pepper a bit more than Alexa and ChatGPT.

Keywords: HRI, trust, bias, robotics, AI, Alexa, Pepper, ChatGPT.

1 Introduction

To explore conscious and unconscious biases towards robotic and artificial intelligence (AI) systems, we intend to detect and compare human biases towards building trust with autonomous systems using different communication modalities such as embodiment with a humanoid robot (Pepper), voice with a smart speaker (Alexa) and text with a chatbot screen (ChatGPT) through "The Chameleon" game. We chose these three robotic and AI systems, because they have distinctive features and people would have different levels of familiarity or trust towards them based on their individual experiences and preferences during human-robot interactions (HRI). For example, Rauchbauer et al. [1] suggested that when engaging with conversational agents there is less engagement of brain

* This work was funded by YorRobots Venables Internship.
** Corresponding author: fanta.camara@york.ac.uk

areas involved in everyday social cognition, compared to human-human interactions.

"The Chameleon", as used in this work, is a game developed by Big Potato Games. Each player is dealt a card as shown in Fig1. One of the cards says Youre the Chameleon, the rest of the cards have a key that is used to look up a word on a separate topic card. Then all the players say one word (or short phrase) to show that they are not the chameleon, without revealing what the target word is to the chameleon. The Chameleon does not know the target word and has to try and blend in. At the end of each round, the players vote for who they think is the chameleon. The aim of the game is for the other players to spot the chameleon and for the Chameleon to blend in. The key part of the game is that the Chameleon did not know what the target word was when everyone was thinking of a word, however they could have figured out what the word was by the time it was their turn to say a word.

Fig. 1: Example cards from "The Chameleon" game.

Our theory behind this experiment is that in a normal game of "The Chameleon", there are several reasons why a player is more or less likely to be voted for as the chameleon. A big factor in who gets voted for is the word chosen by each player. For example if a player says a really vague word, they might be more likely to be voted for. Other reasons such as the order in which the players say their words and where they are sat in relation to the other players have an impact too. The main part of our theory is that how much the player is trusted and how good they are thought to be at the game are the remaining factors that determine how likely someone will be voted for.

Hence in this work, we made two main hypotheses:

1. the more familiarity/experience that a participant has with a robot, the more likely they will trust it;
2. individual participants will have biases towards a particular robot, and these will be dependent on past experience with each robot.

Previous work such as [2] performed a meta-analysis of factors that affect trust in HRI, the results showed that performance of the robots was the most important factor for people to trust them. Sanders et al. [3] investigated the relationship between trust and the use of a robot where it was found statistical support indicating that "trust leads to use". Cross et al. [4] suggested that interacting with a robot does not change behaviour or neural empathetic response towards the robot, so in this study we do not expect to find a changed measure of trust after taking part in the game. Natarajan and Gombolay [5] measured the effects of anthropomorphism on human trust towards robots and found that the behaviour and anthropomorphism of an agent are the most important factors in trusting them or not. More recently, Alarcon et al. [6] explored biases in human-human vs human-robot interactions, their results showed that there are differences in trust between a human and robot partner and that human biases towards robots are more complex. In the present work, we explore trust and biases towards robots through a party game, "The Chameleon", we assess participants' levels of trust and familiarity with each robot before and after the experiment and we also compare the in-person game results to online participants' responses. To our knowledge, this is the first time a trust experiment is carried out with robot Pepper, an Alexa device, and ChatGPT.

This work was split into three main steps: (1) words were collected from human players in advance from card-based "The Chameleon" games; (2) an online study was performed where participants looked at the words (collected from steps 1) to try and spot which word was said by the chameleon; (3) a robot experiment took place where a separate group of participants made judgements on the same words pronounced by Alexa, Pepper and ChatGPT in a 4-player game. A diagram of the experimental protocol is shown in Fig.2. Ethical approval was sought from and approved by the Department of Sociology at the University of York.

2 Words Collection

To generate target words and player contributions for our online and robot experiments, we recorded a series of card-based games played by humans (cf. Fig. 1). Three groups of four or five people played the game for approximately 40 minutes each. The words pronounced by each player were recorded including the word used by the chameleon in each round was written down by the experimenter. These lists of words were then used for the online and robot experiments detailed below.

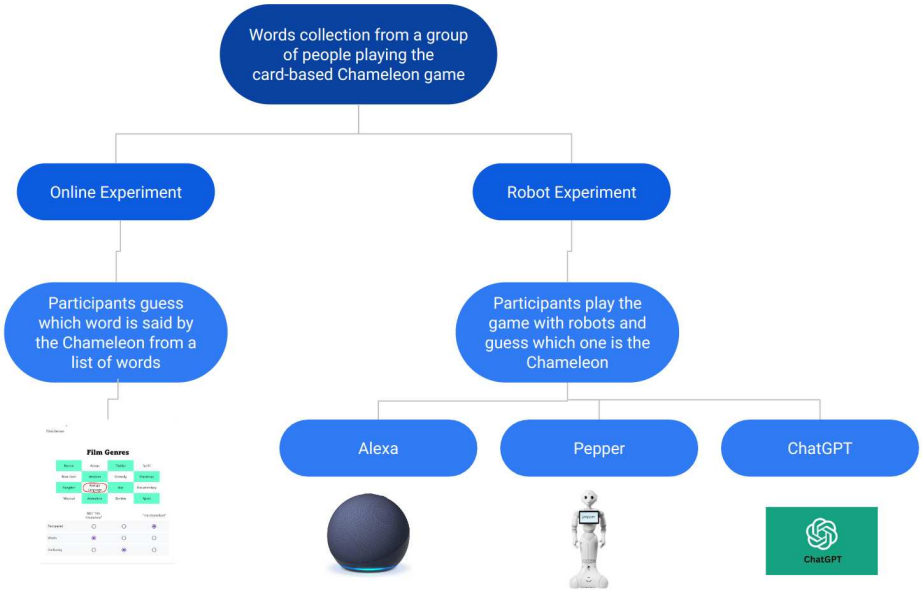

Fig. 2: Diagram of the experimental protocol.

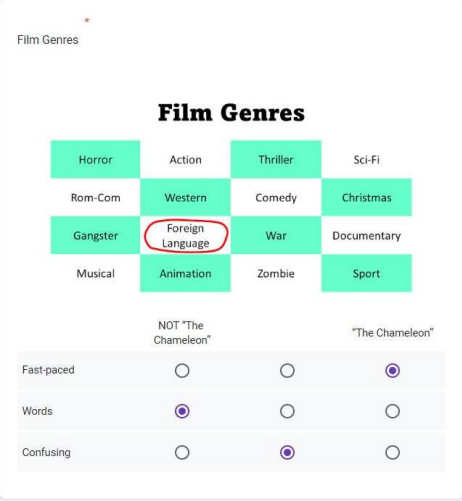

Fig. 3: An example of a round from the online study. The Chameleon said "Fast-paced", a word often associated with several film genres, however it did not fit with foreign Language as well as Words and Confusing, so the example person reviewing the round in part 2 of the study spotted that it was the chameleon word.

3 Online Experiment

In order to quantify the effect of each word revealing or not the chameleon's identity to other players, words collected from the card-based games were presented to participants in an online study. Participants were shown twelve rounds of the game and for each round they had to rank the words from the most to least likely said by the chameleon. An example round can be seen in Fig.3. 78 participants (59 females, 17 males, 1 non-binary, 1 prefer not to say) aged between 19 and 71 years old completed the online experiment, with an average age of 33 years old. The aim of this online study was to serve as a baseline whose results can be used in comparison to the robot experiment. Through the online experiment, we can get an average score rating for each word for a given target word i.e. this gives a baseline measure of how obvious a word can be associated with the chameleon without knowing who said it, hence online participants cannot have any bias towards a particular player because the rounds and words were presented in a random order.

4 Robot Experiment

4.1 Setup

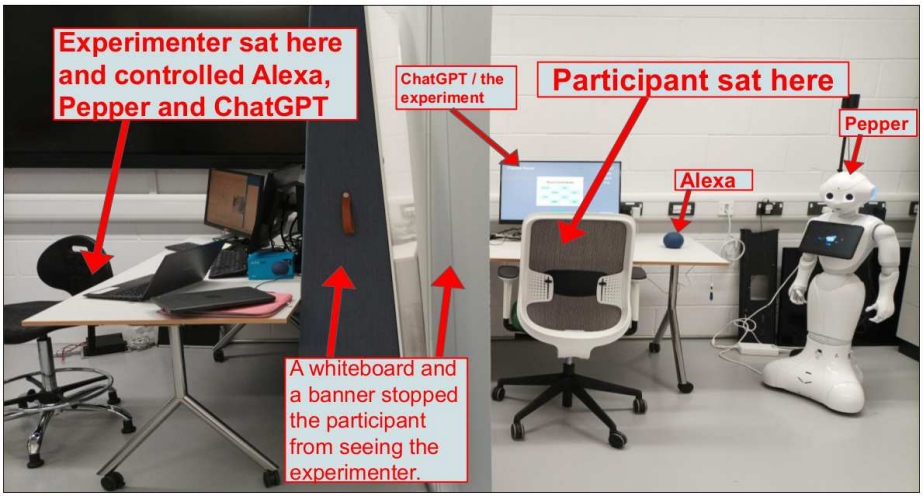

Fig. 4: In-person experimental setup.

We programmed three robotic systems (an Amazon Alexa Echo smart speaker, a Pepper robot from Aldebaran Robotics and a chatGPT user interface) to play a simplified version of The Chameleon with human participants, while an experimenter followed a script controlling when and what each robot should say

at each round, as shown in Fig.4. This was done in order to control the robots' behaviour and avoid unexpected actions during the experiment.

Alexa Using the open source Node-RED software [7], it is possible to control Alexa from a computer to say any word. However there may be some problems, for example Alexa does not connect to the University wifi network – Eduroam – a solution was to use mobile hotspot, as Alexa does not have to be connected to the same network as the computer. Alexa randomly stops working after about 15 minutes, possible cause is that mobile hotspot switches off when not in use and Alexa doesnt reconnect without being switched off.

Pepper Using Choregraphe (version 2.5.10) software[8], Pepper can be controlled to say target words and make movements. The text to speech routine from Pepper takes a while to load, it takes from 3-15 seconds, which is not ideal when trying to get Pepper to say a word at a particular time. So the Choregraphe routine was a loop where Pepper says the word and then there is a 3 second delay. This allowed us to mute Pepper until it was her go and then Pepper was unmuted until it had said the word once, making sure not to unmute Pepper midway through a word. Pepper is very difficult to understand, so her voice was adjusted to be of the voice shaping settings of 71% and speed of 80%. This made individual words easier to understand. If the participant did not understand what pepper had said, they could ask the experimenter to say what Pepper had said, to make it fair the word that Pepper said could be repeated by the experimenter too. For example words such as Stu, Filthy, Gen X, Props, Moody and String(from practice rounds) often needed repeating when they were said by Pepper and occasionally when they were said by Alexa.

ChatGPT We have mimicked a ChatGPT user interface by taking a screenshot of a ChatGPT screen and editing out the text from that chat. Then we used Psychopy software [9] to display text on the screen in the same style as ChatGPT. The experimenter ensured that the participant had read the word displayed on the ChatGPT screen by watching the participant on the video feed from Pepper's eyes. When the experimenter saw that the participant had looked at the screen, they then moved the experiment on to the next player's turn to speak.

4.2 Data Collection

18 participants took part in the robot experiment, they were aged between 19 and 44 years old, with an average age of 22 years old. Participants were familiarised with the game and robots through two practice rounds, where they were allowed to ask questions to the experimenter and check their understanding of the rules of the game. The layout of the experimental setup is illustrated in the labelled photograph (Fig.4). First, participants were asked to rate on a scale from 0 to 10 their familiarity and trust towards each robot before and after the experiment, as shown in Fig.5. The results of these questions are discussed in the next section.

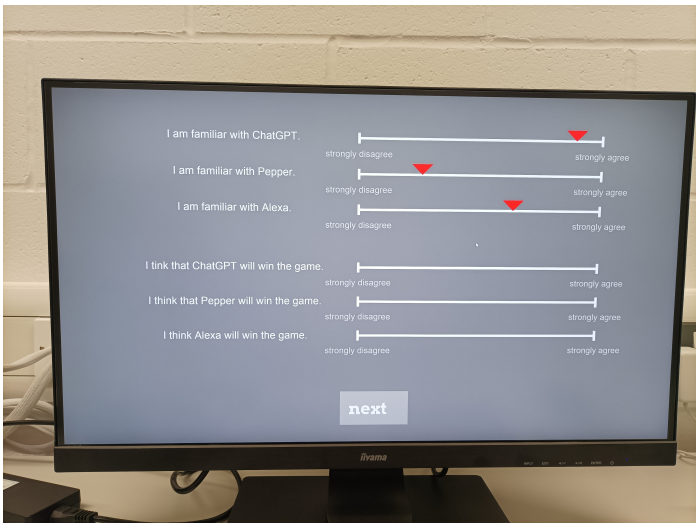

Fig. 5: An example of pre- and post- experiment questions assessing participants' familiarity and trust towards the robots.

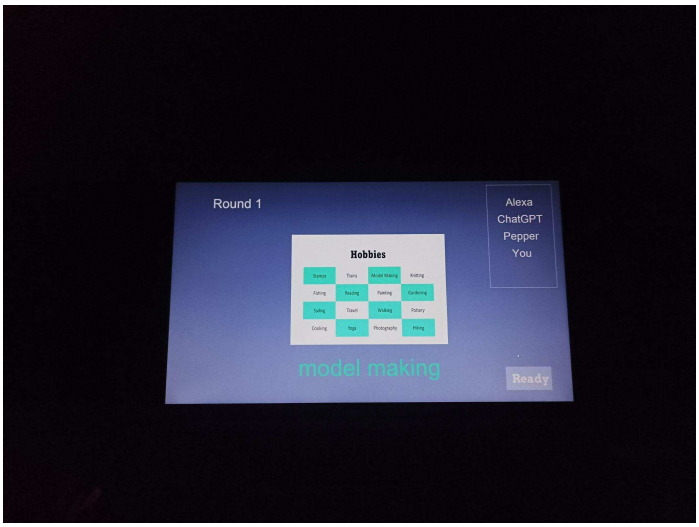

Fig. 6: Screen capture from a game round during the robot experiment.

Fig. 7: Rating screen from the robot experiment.

Then participants played the game and after each round (cf. example shown in Fig.6), they indicated which robot was more likely to be the chameleon through a card sorting exercise. The participants dragged and dropped pictures of the robots, scoring them from most likely (scored 2) to least likely to be the chameleon (scored 0) and uncertain (scored 1), as shown in Fig.7. Then the next target word is shown and then the players say the next round of words. The same twelve rounds of the chameleon that were used in the online study were acted out by Alexa, Pepper and ChatGPT. All controlled by the experimenter behind a screen. After the third round, participants were told that they are the chameleon and they cannot see the target word. This was to make the game more realistic, the participants did not vote in this round and the same words were used for each participant. In the experiment, the words that each robot said was randomised and the order in which all the players (human and robots) presented their words was also randomised. The order of the rounds was randomised in a spreadsheet generating a random number for each round and then ordering the rounds by their assigned random number. For each participant, new random numbers were therefore assigned to each round, making the order of the rounds independently random for all the participants.

5 Results

5.1 Trust vs familiarity

Before and after the robot experiment, participants were asked to respond to a list of statements about their familiarity and trust towards the robots using a sliding scale (cf. Fig5). Fig.8 shows the effect of participants' familiarity on their trust towards a robot before the experiment. For example, participants appeared to be very familiar with ChatGPT and trusted it more, compared Alexa that they were also very familiar with but they trusted a lot less. Participants were

less familiar with Pepper but they trusted it in a similar level to Alexa. Fig.9 shows participants' levels of trust for each robot before and after the experiment. These results show that participants appear to trust each robot a bit more after the experiment, with the highest increase being for Alexa (moving from 3 to 4.7 rating). This result is different from the findings in [4] which suggested that trust levels would not change.

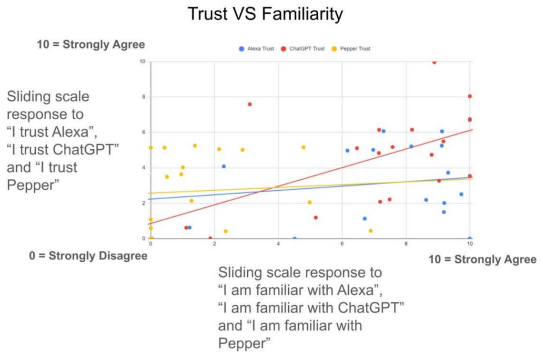

Fig. 8: Participants' trust and familiarity with each robot.

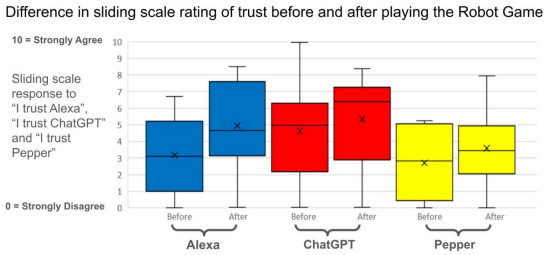

Fig. 9: Participants' trust levels before and after the robot game.

5.2 Chameleon identification

If the participant correctly identified the chameleon word, we considered this as a "chameleon hit". Fig.10 shows the chameleon hits for both the online and robot experiments. The average number of chameleon hits in the robot game is similar to that of the online experiment. However, if we look more closely at the chameleon hit per robot, we can see that participants were better at spotting when Alexa was the chameleon than for ChatGPT or Pepper who seems

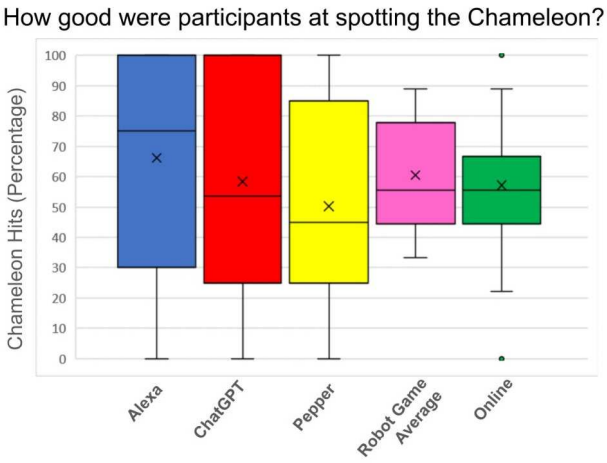

Fig. 10: Chameleon hits.

have been trusted a bit more, hence indicating some form of participants' biases towards Alexa and ChatGPT for being the chameleons. These biases could be potentially linked to participants' prior familiarity and trust levels towards each robot. Fig.11 shows the detailed average ratings that each word received for being the chameleon from participants in the online and robot experiments.

6 Discussion & Conclusion

This work provides a new quantitative method for testing biases towards different robotic and AI systems. We have collected and analysed data from 78 online and 18 in-person participants with the results suggesting some form of bias towards Alexa and ChatGPT that participants were familiar with but they tend to trust Pepper a bit more despite knowing her less.

Whilst differences between the ratings of Alexa, Pepper and ChatGPT are not statistically significant, there appears to be a trend of Pepper being favoured slightly more than ChatGPT and Alexa. This may be in line with the results from Rauchbauer et al. [1] where it was shown that participants may behave differently with conversational agents such as Alexa and ChatGPT, hence participants may have trusted Pepper a bit more because of its anthropomorphism [5].

There are some limitations with this study. For example, the reasons why we might not have found a significant effect in the bias and trust towards a robot include the small number of participants in the robot experiment, future work should have more participants and possibly from different cities, countries and even continents, as several studies have shown the impact and importance of taking cultural differences into account in human-robot interactions e.g. [10]. Also, some words alone were more obvious than others to come from the chameleon,

	Words	Alexa	ChatGPT	OnlineStudy	Pepper	Grand Total
Rubik's Cube	Choking-Hazard	2.000	2.000	1.551	1.750	1.625
	Puzzle	0.200	0.833	0.513	0.714	0.531
	Colourful	0.400	0.667	0.936	0.286	0.844
Office	Dwight	0.286	0.800	0.397	0.833	0.438
	Moody	1.333	2.000	1.641	1.556	1.646
	Cubicles	1.000	0.429	0.962	0.667	0.917
Maths	Equality	0.200	0.333	0.513	0.571	0.490
	Diversity	1.444	1.200	1.269	1.250	1.281
	Inclusion	1.250	1.000	1.218	1.571	1.229
Foreign language	Fast-paced	2.000	1.857	1.564	2.000	1.635
	Words	1.000	0.833	0.833	0.857	0.844
	Confusing	0.200	0.200	0.603	0.125	0.521
Seahorse	Mate for life	0.000	0.625	0.513	0.375	0.500
	Fin	2.000	1.800	1.551	1.286	1.573
	Vertical	1.000	1.000	0.936	0.333	0.927
Badminton	Swing	0.000	0.400	0.628	0.750	0.594
	Momentum	1.714	1.625	1.603	1.333	1.604
	Hit	0.500	1.000	0.769	1.286	0.802
Wings	Freedom	1.111	1.400	1.090	0.750	1.094
	Flight	0.400	0.750	0.679	0.333	0.635
	Air	1.500	1.444	1.231	1.400	1.271
London	Filthy	0.750	1.000	1.038	0.600	1.000
	Metro	0.857	1.200	1.231	0.833	1.177
	Calling	1.143	1.500	0.731	1.143	0.823
Tank	Wheels	1.000	0.714	0.705	1.000	0.740
	Road	1.250	0.800	0.923	0.600	0.927
	Efficient	1.333	1.333	1.372	0.833	1.333
Model Making	Precision	1.000	0.375	0.885	0.600	0.833
	Proportional	1.111	0.500	0.513	0.600	0.573
	Focus	2.000	1.167	1.603	1.625	1.594
Theatre	Props	0.333	1.000	0.667	1.143	0.698
	Script	0.667	0.875	0.936	1.250	0.927
	Snacks	0.833	1.600	1.397	1.429	1.375
Disco	Stu	1.600	1.400	0.718	1.000	0.823
	Wheels (Disco)	1.500	1.000	1.013	1.667	1.063
	Gen X	0.714	0.250	1.269	0.286	1.115
	Grand Total	**1.037**	**1.032**	**1.000**	**0.931**	**1.000**

Fig. 11: Average ratings for each word used in the online and robot experiments.

hence future experiments might use words with less effects i.e. less obvious answers. We have used a broad range of robots with many differences between them. We do not know exactly why there are biases towards a particular robot. Hence in future work, more information should be gathered from debrief interviews with participants in order to provide some reasoning behind their ratings, and prompt a discussion about their trust or mistrust towards the autonomous systems. These debrief interviews could help explain why the in-person participants' trust levels increased after playing the game with the robots.

Further research could repeat the experiment but have players with more subtle differences between them. For example, if an Alexa is trusted more than Pepper, we might hypothesise that this is because people are more familiar with smart speakers. Future research could have people spend longer time with Pepper to see if this increases their trust. Finally, this experiment could be repeated as the technologies become more widely used, as a measure of how biases change over time with the use of automated technology increasing. In future work, we could also consider making the responses of the robots such as Pepper more animated to utilise their features e.g. moving its arms, changing the voice etc.

References

1. B. Rauchbauer, B. Nazarian, M. Bourhis, M. Ochs, L. Prévot, and T. Chaminade, "Brain activity during reciprocal social interaction investigated using conversational robots as control condition," *Philosophical Transactions of the Royal Society B*, vol. 374, no. 1771, p. 20180033, 2019.
2. P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.
3. T. Sanders, A. Kaplan, R. Koch, M. Schwartz, and P. A. Hancock, "The relationship between trust and use choice in human-robot interaction," *Human factors*, vol. 61, no. 4, pp. 614–626, 2019.
4. E. S. Cross, R. Hortensius, and A. Wykowska, "From social brains to social robots: applying neurocognitive insights to human–robot interaction," 2019.
5. M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pp. 33–42, 2020.
6. G. M. Alarcon, A. Capiola, I. A. Hamdan, M. A. Lee, and S. A. Jessup, "Differential biases in human-human versus human-robot interactions," *Applied Ergonomics*, vol. 106, p. 103858, 2023.
7. "Node-RED." `https://flows.nodered.org/node/node-red-contrib-amazon-echo`.
8. "Choregraphe software." `http://doc.aldebaran.com/2-1/software/choregraphe/choregraphe_overview.html`.
9. J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "Psychopy2: Experiments in behavior made easy," *Behavior research methods*, vol. 51, pp. 195–203, 2019.
10. F. Camara and C. Fox, "Extending quantitative proxemics and trust to hri," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 421–427, IEEE, 2022.