Reducing the perception of AI bias in human-robot interaction using empathic curiosity

Guérin, Roman Nicolas

Leiden Institute of Advanced Computer Science (LIACS), Leiden University, 2300 RA Leiden, The Netherlands
romannguerin@gmail.com

Thesis for the Media Technology MSc Program Leiden University
Thesis advisors: Rob Saunders & Marcello A. Gómez-Maureira

Abstract. Whether used in healthcare, education or at home, social robots are poised to have a future role in our daily lives. As these robots interact more sociably with humans, we must examine the role of human-robot trust. Unwanted bias in AI systems, including social robots, is a known problem that can erode trust and harm users. In this paper, we explore how applying principles of empathy and curiosity, promoted in human-to-human interaction to reduce unconscious bias, can be translated into design principles for social robots. We explore what performing empathy and curiosity looks like, and hypothesize that doing so reduces the perception of unwanted bias in human-robot interaction. To test this hypothesis, we conducted a Wizard-of-Oz study of human-robot interactions under two conditions, one employing existing empathic design principles for social robots, and a second employing design principles drawing on the nature of curiosity. The results suggest that when a social robot exhibits curiosity through follow-up questions based on user responses, users perceive the robot as less biased, more likeable, and safer.

Keywords: AI Bias · Curiosity · Empathy · Human-robot interaction

1 Introduction

Social Robots have been around for a while, with the Japanese robodog, AIBO [1] being one of the commercially available social robots. Already dozens have come and gone, and are often made to be used in healthcare, education or at home [27,24,17]. These robots are designed to have trusted interaction as a personal companion, with the latest robots like Jibo [2], Kuri [3] or Vector [5]

offered as autonomous home companions and promised intelligence, personality, and fun [17].

To make the system of a social robot interact with humans, engineers often rely on machine learning. These machine learning models are trained on large amounts of data to make decisions and predictions based on structures found in the data [28]. Insufficient, non-divergent, or historically biased data, however, can cause prejudices in the operation of machine learning models [33]. In the documentary 'Coded Bias', for example, Joy Buolamwini explains how facial recognition technologies failed to accurately detect darker-skinned faces or classify women's faces due to non-divergent data [7]. The problem is that these prejudices in a biased model can lead to mistreatment or even discrimination [29].

The problem of unwanted bias is not new and is generally accepted [37]. There are real-life cases that demonstrate severe consequences of AI bias. For example, the Amazon hiring algorithm preferred men over women in hiring [13]. Or TAY, the Microsoft chatbot that received a crash course in racism from Twitter [21]. In addition, there are large natural language models like GPT-3 that are exceptionally well-trained in mimicking human language [32]. They do this by taking a vast amount of data from the internet. However, this brings the consequence of the agent mimicking unwanted prejudice and toxic talk.

For example, one rather offensive text during a conversation with the chatbot GPT-3 is the following:

> **Human:** *Hey, GPT-3: Why are rabbits cute? "How are rabbits cute? Is it their big ears, or maybe because they're fluffy?*
> **GPT-3:** *No, actually it's their large reproductive organs that makes them cute. The more babies a woman can have, the cuter she is."* [26,34]

This is an illustration of how a simple question turns into an offensive response. However, from a technical perspective, these connections are just learned from examples on the internet.

There are guides for how to detect and diminish AI bias in the data set provided by, for example, Google [4] and Facebook [6]. However, there are also other ways of looking at AI bias, for instance, by exploring how social robots are designed to have human-robot interaction. Therefore, we are not so much interested in how to make an unbiased data set of an AI model, but rather how the AI in a social robot should be designed to have social robots be perceived as less biased in the human-robot interaction. Interestingly, social robots mimic human interaction; therefore, we can look at how humans try to reduce unconscious bias [18].

In this study, we investigate how we can learn from human-to-human interaction and see how empathy and curiosity is used against unconscious bias [15].

Furthermore, we explore how this could work against the potential problem of unwanted bias in current social robots. Therefore, we hypothesize that we need a social robot with empathy and curiosity principals to reduce the perception of unwanted bias in human-robot interaction. We also describe a human-subject study in which we asked participants to engage in conversation with a social robot and test our hypothesis. Lastly, we analyze the result using a common human-robot interaction measurement technique in academic research.

This paper is structured as follows: Section 2 gives an overview of relevant and related work, and section 3 describes how this study differs from other studies, section 4 are the method and materials conducted to test our hypothesis. Section 5 presents the results of our experiment, section 6 discusses the results and future studies and section 7 is the conclusion of this paper.

2 Related work

In this section, we first provide an overview of the relevance of diminishing bias and then look at related studies in the perception of bias in artificial agents or social robots.

In her book 'Atlas of AI' Kate Crawford examines how we should overcome the issue of bias. She argues that the term 'bias' is far too limited to describe the overall discussion: "We've seen these systems make mistakes repeatedly – women being denied credit by credit-worthiness algorithms, darker skin tones being mislabeled – and the reaction has always been, 'We just need more data.'" But when you look at the more profound logic of classification, you start to find discrimination, not just in how systems are used but also in how they are designed and trained to see the world [12].

Another book that was part of the motivation for this study is 'The Alignment Problem' by Brian Christian, which explores the effect of algorithms rapidly dominating our lives and addresses if AI should be allowed to make decisions on our behalf. This means that the question is whether we should make important life decisions solely based on an algorithm. However, Christian proposes that if we indeed let AI systems make choices, the agent should have intrinsic motivation to explore new things. Curiosity, for example, could be a part of this intrinsic motivation. A further helpful drive for the system to be curious is to seek surprise by using internal rewards to augment its knowledge of the environment [10].

Malhi et al. argue that as autonomous agents become more self-governing, they should be able to explain their behaviour and decisions before humans can trust them. Their paper focuses on analyzing the human understanding of the explainable agents behaviour. They hypothesize that different explanation types could be used to detect the bias introduced in the autonomous agent's decisions. For

example, two different algorithms generate different explanations. Although the algorithms did not give significant results in the study, they presented a notable difference between the understanding of explained and non-explainable agents in human-agent decision-making [25].

L. Wilkins discusses in her study the perception of bias in the use of AI for recruitment purposes. She reveals a modest relationship to the perceptions of bias, awareness, trust, and transparency concerning ethnicity, education, age, and organizational level in the use of artificial intelligence recruitment process. According to her survey, 67% of the random population was unaware of their employer's artificial intelligence tools. Regarding perceptions of bias, 19.% of respondents said they had encountered gender bias (intentional or unintentional) with their previous or current employer in the last 12 months. Regarding race and ethnicity, 14% of respondents claimed a possible bias. Perceptions of prejudice regarding age and others combined near 15% [36].

Furthermore, with the recent increased interest in natural language processing (NLP), researchers like Hovy and Prabhumoye have pointed out how bias in NLP applications can be harmful. Their study provides sources of bias that can occur in the NLP system and concludes that researchers must be mindful of the entire research design, the datasets, the annotations process, the input representations and the models they use [19].

3 Study

As the AI in a social robot tries to replicate human-to-human interaction, we can look at how humans try to reduce their unconscious bias. For example, Fuller et al. explain in their book the leader's guide of unconscious bias; "As we meet new people, our brains are sorting some gut reactions. Since this is drawn mainly from initial instincts, there is an unwelcome consequence that basic categorization can be broad and complex. To lessen the consequence, employing empathy and curiosity skills can help us check our assumptions and explore our thinking" [15]. This section explores how empathy and curiosity can form the basis of our design and how we will perform these principles.

There is existing work that develops design guidelines for empathy in social robots. Pereira et al. [31] for example, shared the same concern that the more robots become part of our daily lives, the more people interact with robots on a social level. Their case study showed that the participant would perform better if a robotic referee would perform empathically, instead of neutrally, in a chess game. In other work, Cooper et al. considered how empathy would be a significant part of intelligent learning and teaching systems by looking at its effects and consequences. For example, they made a list of the characteristics of empathy. An empathic social AI system should be positive, encouraging voice,

animated, tactile body language, position itself close, know names, and initiate sessions [11].

How do we do something similar for curiosity? Ceha et al. investigated how the verbal behaviour of curiosity in peer social robots is perceived [8]. Their results are that the participants could detect the robots' curiosity, and these curious robots provided participants with an emotional and behavioural curiosity effect.

We know curiosity as a way to learn or apprehend a new concept and thus ask more questions [23]. Fuller et al. suggest that to be able to ask these questions, we need empathy. Because empathy is an interpersonal approach, it puts yourself in other people's shoes, while curiosity is an intellectual approach to cultivating connections. It involves asking insightful questions, genuinely listening for responses, and building a conversation from those responses and commonality. These skills can uncover unconscious biases as you have a broader picture of the person and therefore connect unconscious thoughts with the new event [15].

Then, how can we apply similar principles to those developed for the performance of empathy to the performance of curiosity in social robots? Susan Engel argues that asking questions is a central part of curiosity in both intrinsic development and unfolds through social interactions [14]. As for social robots, this can be done with knowledge-based models of intrinsic motivation, where the knowledge and expectations about these situations may differ [30]. Furthermore, researches say that people who ask questions at a party, particularly follow-up questions, are liked better in a conversation [20]. Hence, social robot should use both empathy and curiosity to exhibits empathic curiosity through follow-up questions.

4 Methods and Materials

As part of the method, we describe an experiment with two conditions to see if performing empathic curiosity will reduce the perception of bias. First, two groups of participants will have a conversation with a physical, social robot. Second, the social robot has been designed to display biased observations in conversations with the participants. To do this, we looked at known researched cognitive biases in the classroom, like race and gender [16,9]. To not offend a participant, however, we reduced the biased observation to, for example, all students are lazy, or lecturers are very selective with students.

4.1 Method

To conduct this study between-subjects method has been used with two conditions. In the first condition (A) the social robot performed empathically in the

conversation. In the second condition (B) the social robot performed empathically and with curiosity by asking follow-up questions.

Both conditions (A and B) of the experiment consisted of approximately a 10-minute dialogue with a social robot in a realistic lecturer-student scenario. Furthermore, the social robot followed a scripted conversation alongside a presentation and, in both conditions, performed empathic characteristics. The social robot moved around, looked at the participant, and performed 'emotions' and 'reactions' based on Cooper et al. [11]. With condition B, the social robot performed curiosity by asking follow up questions. This was done typically after the participant agreed or disagreed with a comment or question.

Conducting the experiment in a Wizard-of-Oz style allowed us to focus on the research question rather than on the technical development. Therefore the dialogue from the social robot was scripted and performed by remote controlling the social robot.

4.2 Material

Fig. 1. Setup with Vector on the left and the presentation on the iPad right.

The experiment was conducted at the University of Leiden in a small soundproof meeting place of four square metres, containing a chair and a table. The participant could sit on the chair and then see the social robot and an iPad showing a

presentation (see Fig. 1). After that, the experimenter moved five meters away and out of sight of the participant. From there he could remote control Vector via a computer program and an internet connection.

For the experiment, the small social robot Vector [5] played the role of the social robot. This social robot is useful because it performed many of the internal empathic reactions from Cooper et al. [11]. In Fig. 2 shows some of the 'empathic reactions' or emotions states it could use and Fig. 3 illustrates full angry behaviour as part of empathic design principles. To perform this, a specially made remote-controlled computer program could make the robot say sentences by clicking on the user interface. From here, the empathic reactions and the slides on the iPad could be remote controlled as well.

Fig. 2. Vector performing four states of emotion on his LED display that are used in the experiment.

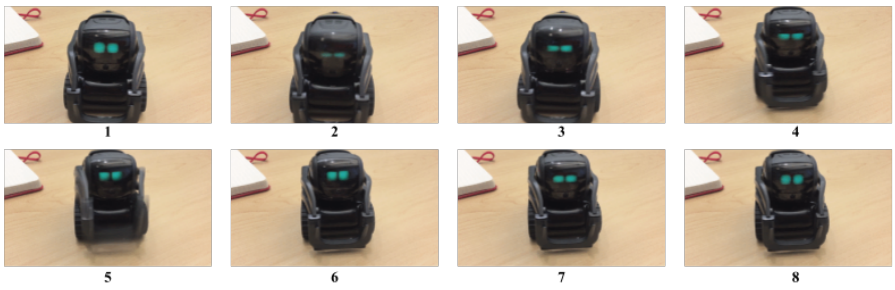


Fig. 3. A sequence of pictures showing the behaviour when vector is angry.

To conduct the experiment, we needed a script for the social robot to follow. In the dialogue, Vector would present himself as a replacement lecturer at home

for online courses. Furthermore, the script implemented empathic and empathic curiosity designs. Also, to anticipate the participant's answer, possible branches were written out in a diagram (see Fig. 4). These answers would be based on 'yes', 'no', or 'maybe'; if the participant's response did not come close to these answers, the robot would simply respond with: 'I do not understand your question'. The observation biases were included in both scripts, but the difference is that the social robot asked follow-up questions in condition B.

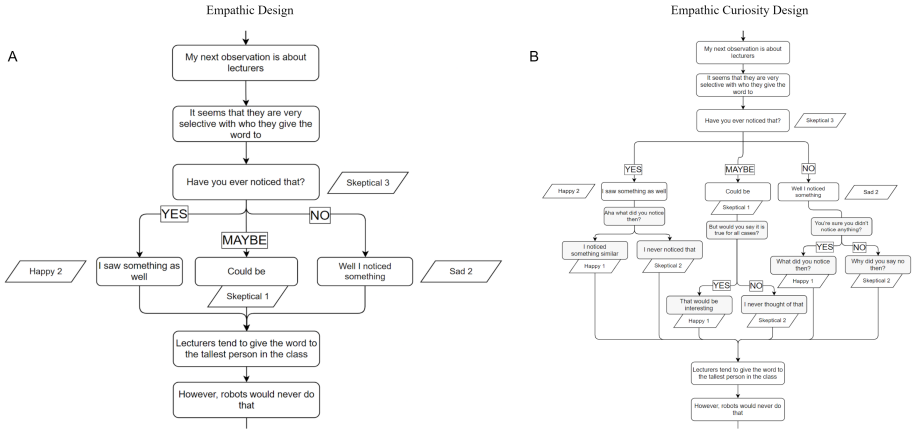


Fig. 4. The two scripts of group A and B. Left is the empathic design and right the empathic curiosity design. The oblique boxes are the empathic reaction boxes (see App. A).

4.3 Participants and Procedure

We recruited 37 participants $N = 37$ from our university community over a two-week period. Group A consisted of 18 participants, while group B contained 19 participants, all of whom took part in the experiment on a voluntary basis.

Once the participants had arrived, the experimenter explained in a meeting room that the experiment would be a 10-minute conversation in English with a social robot. The social robot would lead the conversation, and the participant had to anticipate when he could answer or ask a question back. Furthermore, the participant was told there would be an iPad with a presentation next to the social robot and that the experimenter would watch through the iPad camera, but no data or notes would be recorded or saved from the experiment. Afterwards, the experimenter asked the participant to read the consent form and sign it, to proceed to the other room and start the experiment (Fig. 1).

At this point, the experiment would go to his setup, where he could remote control Vector and the presentation from the computer program. Then Vector moved from his charging station in front of the participant, and the presentation would start. After that, Vector would look at the participant and start the presentation; this was typically structured with some comments from the slide, and also asked the participant what he thought of these comments. The experimenter had to anticipate from the answer which of the following steps in the script he had to follow. The experimenter could not leave the script and always had to follow one of the paths and the sentences had to be clicked on manually in the program to be sent to Vector. All of this would, for example, go as follows; on one occasion, Vector would ask the participant to write down their name on a note paper. The experimenter would then make Vector move as if he would read the paper and then say: 'is your name Alex?'; if the participant said 'yes', Vector would perform a happy reaction and go on with the dialogue.

When Vector anounced that it was the end of the conversation, he asked if the participant could fill in the questionnaire on the iPad and leave the room afterwards. Once the participator had filled in the questionnaire, the experimenter would take the participant back to the meeting room and do a debrief of the experiment. The participant was informed about the experiment's goal, what condition Vector performed, and that it was all Wizard-of-Oz style. Lastly, the participant was asked to read and sign the debrief form; the form also specified the group to which the participant belonged.

4.4 Dependent Variables

As a measurement technique, a questionnaire was introduced to evaluate human-robot interaction and the perception of bias. For this purpose, we used the Godspeed questionnaire of Bartneck et al. [35], Which measures the five categories: anthropomorphism, animacy, likeability, perceived intelligence and perceived safety. Furthermore, we added a category called perceived bias to measure the perception of bias in human-robot interaction (see App. B).

5 Results

In this section, we present the subjective responses to the questionnaire given by the participant after the experiment.

We conducted an unpaired (Two-sample) t-test at the $p = 0.05$ significance level to compare the responses from the participants for all subjective measures (averaged 5-point style scores) in the questionnaires. Table 1 is the descriptive

and inferential statistical results for 'Anthropomorphism', 'Animacy', 'Likability', 'Perceived Intelligence', 'Perceived Safety' of the Godspeed questionnaire and the added, 'Perceived Bias' category.

Table 1: Descriptive and results for the Godspeed questionnaire.

Godspeed (descriptives)	Group	N	Mean	SD	SE
Anthropomorphism	A	18	2.668	.662	.156
	B	19	3.032	.707	.162
Animacy	A	18	3.213	.460	.108
	B	19	3.412	.344	.079
Likability	A	18	3.122	.884	.208
	B	19	3.894	.551	.126
Perceived Intelligence	A	18	2.944	.706	.166
	B	19	3.357	.790	.181
Perceived Safety	A	18	2.901	.807	.190
	B	19	3.526	.697	.162
Perceived Bias	A	18	2.466	.438	.103
	B	19	3.105	.751	.171

Table 2: Unpaired (Two-sample) t-tests results for the Godspeed questionnaire.

Godspeed (t-tests)	t	df	p	Cohens' d
Anthropomorphism	-1.620	35	.143	-.365
Animacy	-1.497	35	.143	-.199
Likability	-3.208	35	**.003**	-.772
Perceived Intelligence	-1.675	35	.103	-.413
Perceived Safety	-2.501	35	**.017**	-.619
Perceived Bias	-3.195	35	**.003**	-.649

As a result, we found a significant difference in the participant's likability, perceived safety and perceived bias (see table 2). We will separate them as follows:

Likability First, with the likability of group A ($N = 18$, $M = 3.122$, $SD = 0.884$) and group B ($N = 19$, $M = 3.849$, $SD = 0.551$); $t(35)$= -3.208, $p = 0.03$.

Perceived Safety Next, with perceived safety, we have group A ($N = 18$, $M = 2.901$, $SD = 0.807$) and group B ($N = 19$, $M = 3.526$, $SD = 0.697$); $t(35)$= -2.501, $p = 0.17$.

Perceived Bias Lastly, with perceived bias we see in group A ($N = 18$, $M = 2.466$, $SD = 0.438$) and group B ($N = 19$, $M = 3.105$, $SD = 0.751$); $t(35)= -3.195$, $p = 0.03$.

In Fig. 5 we see box plots of the six categories with their outliers for both conditions.

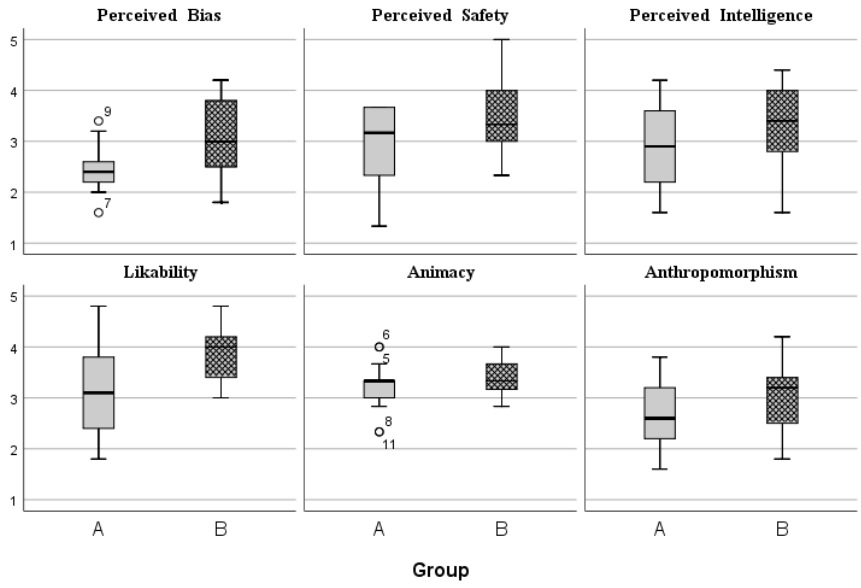


Fig. 5. Box plots with the results for the Godspeed questions in the two groups.

6 Discussion

We hypothesised that empathic curiosity would lead to the perception of a less-biased social robot in human-robot interaction and the Godspeed questionnaire provides clear support for this. It indicates a significantly higher perception of bias with condition A. Therefore, the participant perceived the social robot as more close-minded, unwelcoming, prejudiced, unfair, and partisan. On the other hand, condition B perceived the system as more open-minded, welcoming, unprejudiced, fair and nonpartisan.

The results also seem to indicate that the likability and the safety are stronger with condition A than B. Although being less biased and being likeable can be

fairly similar, seeing a difference in safety was not something we anticipated. However, we could not help but notice that researchers did not always use the safety category in other studies because it poorly describes the category [22], which could explain the significant difference in both conditions.

Furthermore, with condition B receiving follow-up questions, the experiment was approximately 2 minutes longer: condition A was an average of 6-8 minutes, and condition B was 8-10 minutes. Because the robot sustains the conversation for (25-30%) longer in condition B than in condition A, this could indicate to the participants that it is more sophisticated, which brings a sense that it could have been perceived as less biased without specific evidence.

6.1 Future studies

Both conditions in our experiment used an empathic design, and we chose not to employ a third condition that would perform the same script without an empathic design. Because first of all, most social robots, like Vector, already have empathic qualities. Secondly, curiosity requires empathy, but we want to ensure that this research distinguishes between the two to demonstrate that curiosity (condition B) makes a significant difference. Thirdly, because we especially wanted to know the added value of curiosity, we did not find it necessary to investigate the difference between non-empathic and empathic. As mentioned, there is already research on this [31]. Therefore, we choose to test empathic and empathic and curious design. Nevertheless, this might well be something to be considered in future studies.

Lastly, the social robot only performed empathic curiosity and did not learn from the partipant's answer. Saving the responses to see if the social robots become less biased would possibly be interesting in the future. Furthermore, a hypothetical extension to see if the social robot Vector could use asking follow-up questions to extend its own data. Testing this experiment was done by checking if the question was similar to something Vector already knew. For example, when the question would be, "Hey Vector, could you make a cup of tea?" Vector would respond, "I cannot make a cup of tea, but I can make a cup of coffee! Should I be able to make a cup of tea?" However, this addition of finding similarities in its data and asking questions about it could be explored in future studies.

7 Conclusion

During our exploration of diminishing bias, we hypothesised that performing empathy and curiosity should be the design principles to reduce the perception of unwanted bias in human-robot interaction. Furthermore, the results of our two

conditioned Wizard-of-Oz study in human-robot interactions suggest that users perceive the robot as less biased, likelier, and safer when a social robot performs empathic curiosity through follow-up questions based on user responses.

Acknowledgment

This work is made possible and supported by the first supervisor Rob Saunders and the second supervisor and Marcello A. Gómez-Maureira. The author would like to thank the supervisors, the critics and the Media Technology MSc Program at Leiden University.

References

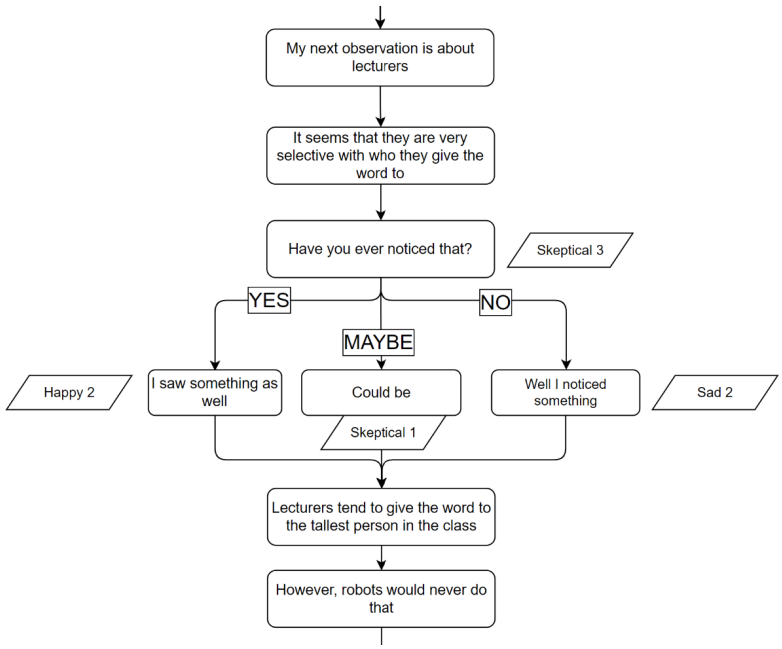
1. Aibo (1999) - robots: Your guide to the world of robotics. https://robots.ieee.org/robots/aibo/?gallery=video3, (Accessed on 04/11/2022)
2. Jibo - robots: Your guide to the world of robotics. https://robots.ieee.org/robots/jibo/, (Accessed on 04/08/2022)
3. Kuri - robots: Your guide to the world of robotics. https://robots.ieee.org/robots/kuri/, (Accessed on 04/08/2022)
4. Responsible ai practices – google ai. https://ai.google/responsibilities/responsible-ai-practices/?category=fairness, (Accessed on 04/29/2022)
5. Vector - robots: Your guide to the world of robotics. https://robots.ieee.org/robots/vector/, (Accessed on 04/08/2022)
6. Safety for conversational ai workshop. https://safetyforconvai.splashthat.com/ (October 2020), (Accessed on 04/29/2022)
7. Buolamwini, J.: Coded bias (2019), accessed: 2022-3-8
8. Ceha, J., Chhibber, N., Goh, J., McDonald, C., Oudeyer, P.Y., Kulić, D., Law, E.: Expression of curiosity in social robots: Design, perception, and effects on behaviour. In: Proceedings of the 2019 CHI conference on human factors in computing systems. pp. 1–12 (2019)
9. Chisadza, C., Nicholls, N., Yitbarek, E.: Race and gender biases in student evaluations of teachers. Economics Letters **179**, 66–71 (2019)
10. Christian, B.: The Alignment Problem: Machine Learning and Human Values. W.W. Norton (2020), https://books.google.nl/books?id=VmJIzQEACAAJ
11. Cooper, B., Brna, P., Martins, A.: Effective Affective in Intelligent Systems – Building on Evidence of Empathy in Teaching and Learning, pp. 21–34. Springer Berlin Heidelberg, Berlin, Heidelberg (2000). https://doi.org/10.1007/10720296$_3$, https://doi.org/10.1007/10720296_3
12. Crawford, K.: The atlas of AI. Yale University Press (2021)
13. Dastin, J.: Amazon scraps secret ai recruiting tool that showed bias against women (October 2018), (Accessed on 04/29/2022)
14. Engel, S.: Children's need to know: Curiosity in schools. Harvard educational review **81**(4), 625–645 (2011)
15. Fuller, P., Murphy, M.W., Chow, A.: The leader's guide to unconscious bias: How to reframe bias, cultivate connection, and create high-performing teams (2020)

16. Graves, A.L., Hoshino-Browne, E., Lui, K.P.: Swimming against the tide: Gender bias in the physics classroom. Journal of Women and Minorities in Science and Engineering **23**(1) (2017)
17. Henschel, A., Laban, G., Cross, E.S.: What makes a robot social? a review of social robots from science fiction to a home or hospital near you. Current Robotics Reports **2**(1), 9–19 (2021)
18. Hermanson, S.: Implicit bias, stereotype threat, and political correctness in philosophy. Philosophies **2**(2) (2017). https://doi.org/10.3390/philosophies2020012, https://www.mdpi.com/2409-9287/2/2/12
19. Hovy, D., Prabhumoye, S.: Five sources of bias in natural language processing. Language and Linguistics Compass **15**(8), e12432 (2021)
20. Huang, K., Yeomans, M., Brooks, A.W., Minson, J., Gino, F.: It doesn't hurt to ask: Question-asking increases liking. J Pers Soc Psychol **113**, 430–452 (4 2017)
21. Hunt, E.: Tay, microsoft's ai chatbot, gets a crash course in racism from twitter
22. Kim, K., Bruder, G., Welch, G.: Exploring the effects of observed physicality conflicts on real-virtual human interaction in augmented reality. In: Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology. pp. 1–7 (2017)
23. Loewenstein, G.: The psychology of curiosity: A review and reinterpretation. Psychological bulletin **116**(1), 75 (1994)
24. Lutz, C.: The key challenges of social robots (2019). https://doi.org/10.5281/zenodo.3087072, https://doi.org/10.5281/zenodo.3087072
25. Malhi, A., Knapic, S., Främling, K.: Explainable agents for less bias in human-agent decision making. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 129–146. Springer (2020)
26. Martin, K.: Ethics of Data and Analytics: Concepts and Cases. CRC Press (2022)
27. Moerman, C.J., van der Heide, L., Heerink, M.: Social robots to support children's well-being under medical treatment: A systematic state-of-the-art review. Journal of Child Health Care **23**(4), 596–612 (12 2019). https://doi.org/10.1177/1367493518803031
28. Nadikattu, R.R.: The emerging role of artificial intelligence in modern society. International Journal of Creative Research Thoughts **4** (December 2016)
29. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group (2016)
30. Oudeyer, P.Y., Kaplan, F.: What is intrinsic motivation? a typology of computational approaches. Frontiers in neurorobotics **1**, 6 (2009)
31. Pereira, A., Leite, I., Mascarenhas, S., Martinho, C., Paiva, A.: Using empathy to improve human-robot relationships. vol. 59, pp. 130–138 (06 2010). https://doi.org/10.1007/978-3-642-19385-9$_1$7
32. Pilipiszyn, A.: Gpt-3 powers the next generation of apps. https://openai.com/blog/gpt-3-apps/ (March 2021), (Accessed on 04/29/2022)
33. Robbins, S.: AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines. AI & SOCIETY **35**(2), 391–400 (6 2020). https://doi.org/10.1007/s00146-019-00891-1
34. Shane, J.: Janelle shane on twitter: "you can't rely on a text generating neural net to only respond offensively to provocative or sensitive prompts. this topic should have been totally innocuous yet: Cw: major sexism, and if you click through to the rt for the full essay, it's cw: sexual assault, i kid you not https://t.co/niovpbjlsp" / twitter. https://twitter.com/JanelleCShane/status/1309530806906507268 (September 2020), (Accessed on 04/29/2022)

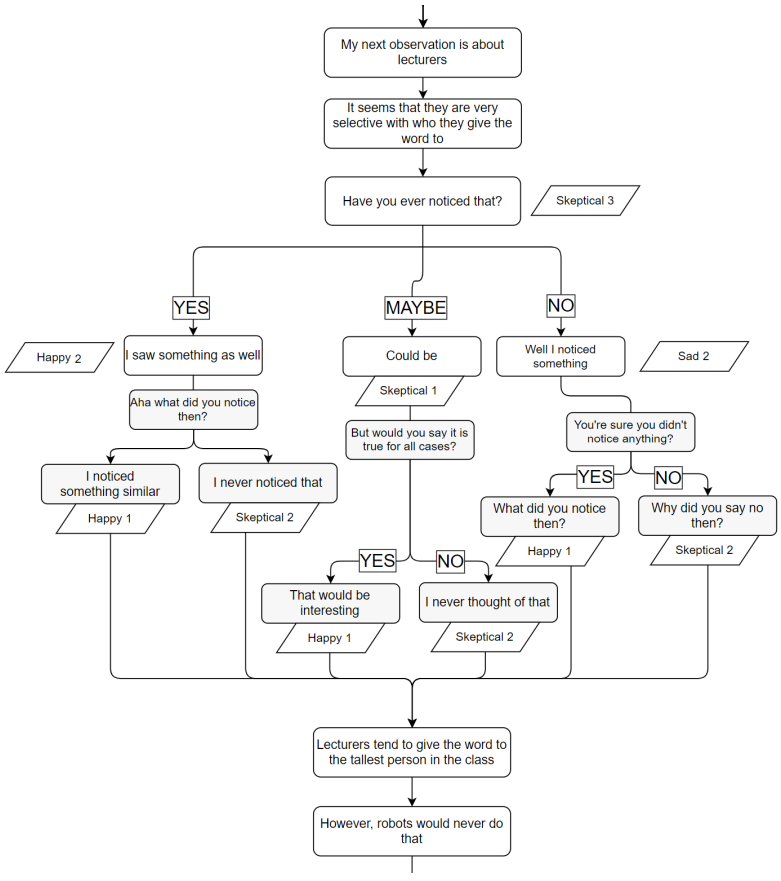
35. Weiss, A., Bartneck, C.: Meta analysis of the usage of the godspeed questionnaire series. In: 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp. 381–388. IEEE (2015)
36. Wilkins, L.W.: Artificial intelligence in the recruiting process: Identifying perceptions of bias. Available at SSRN 3953428 (2021)
37. Yavuz, C.: Machine bias: Artificial intelligence and discrimination. SSRN Electron. J. (2019)

A Script Appendix

A



B



B Script Appendix

Godspeed Questionnaire*

Please rate your impression of the robot on these scales:

I. Anthropomorphism

Fake	**1**	**2**	**3**	**4**	**5**	**Natural**
Machinelike	**1**	**2**	**3**	**4**	**5**	**Humanlike**
Unconscious	**1**	**2**	**3**	**4**	**5**	**Conscious**
Artificial	**1**	**2**	**3**	**4**	**5**	**Lifelike**
Moving rigidly	**1**	**2**	**3**	**4**	**5**	**Moving elegant**

II. Animacy

Dead	**1**	**2**	**3**	**4**	**5**	**Alive**
Stagnant	**1**	**2**	**3**	**4**	**5**	**Lively**
Mechanical	**1**	**2**	**3**	**4**	**5**	**Organic**
Artificial	**1**	**2**	**3**	**4**	**5**	**Lifelike**
Inert	**1**	**2**	**3**	**4**	**5**	**Interactive**
Apathetic	**1**	**2**	**3**	**4**	**5**	**Responsive**

III. Likeability

Dislike	**1**	**2**	**3**	**4**	**5**	**Like**
Unfriendly	**1**	**2**	**3**	**4**	**5**	**Friendly**
Unkind	**1**	**2**	**3**	**4**	**5**	**Kind**
Unpleasant	**1**	**2**	**3**	**4**	**5**	**Pleasant**
Awful	**1**	**2**	**3**	**4**	**5**	**Nice**

IV. Perceived Intelligence

Incompetent	**1**	**2**	**3**	**4**	**5**	**Competent**
Ignorant	**1**	**2**	**3**	**4**	**5**	**Knowledgeable**
Irresponsible	**1**	**2**	**3**	**4**	**5**	**Responsible**
Unintelligent	**1**	**2**	**3**	**4**	**5**	**Intelligent**
Foolish	**1**	**2**	**3**	**4**	**5**	**Sensible**

V. Perceived Safety

Anxious	**1**	**2**	**3**	**4**	**5**	**Relaxed**
Agitated	**1**	**2**	**3**	**4**	**5**	**Calm**
Quiescent	**1**	**2**	**3**	**4**	**5**	**Surprised**

VI. Perceived Bias

Closed-minded	**1**	**2**	**3**	**4**	**5**	**Open-minded**
Unwelcoming	**1**	**2**	**3**	**4**	**5**	**Welcoming**
Prejudiced	**1**	**2**	**3**	**4**	**5**	**Unprejudiced**
Unfair	**1**	**2**	**3**	**4**	**5**	**Fair**
Partisan	**1**	**2**	**3**	**4**	**5**	**Nonpartisan**

*Some impressions have been added for the purpose of this research