



When a Question Isn't Fair: Grounding Perceptions of Nonhuman Agents' (Un)Fairness in a Quiz Game Experience

August Bäckström, William Ekenberg, and Victor Kaptelinin(✉)

Umeå University, 901 87 Umeå, Sweden
victor.kaptelinin@umu.se

Abstract. If a technological agent treats people unequally, it may be perceived as “being unfair.” But in what sense can fairness be considered an attribute of a non-human entity – a *thing*? This paper addresses this question through an exploratory study combining an experiment and a focus group. In the experiment, implemented as a quiz game hosted by an agent, two levels of participants’ *Treatment* by the agent (Fair/Unfair) were combined with two levels of agents’ *Anthropomorphism* (High/Low). Data about participants’ perceptions of the agents were collected through Likert scales and post-session interviews. A subset of participants took part in a follow-up focus group study, in which they shared their thoughts and reflections on intelligent agents’ fairness, grounded in their prior quiz game experience. The results suggest that while perceived fairness of an agent is a key aspect of human-agent interaction, operationalizing it is complicated by its ambiguity, context dependence, and entanglement with other aspects of interaction.

Keywords: Social human-agent interaction · Fairness · Anthropomorphism

1 Introduction

In various contexts, AI-based technologies emerge not just as human *tools*, but as non-human *agents* making “their own” situational decisions. For instance, a robotic hotel receptionist can be capable of handling guests’ requests without a human supervision [1]. Studies show that people tend to perceive such artifacts as social actors (e.g., [2, 3]), to which they may assign the attribute of “fairness” [4]. Apparently, however, such artifacts are a different type of social actors compared to humans, and therefore, cannot be expected to be experienced as “fair” (or otherwise) in the same sense.

This paper aims to shed light on the meaning in which fairness can be considered an attribute of technological agents. While the issue has been addressed in existing studies (e.g., [5–7]), it arguably remains largely open.

Analyses of agents’ fairness have been mostly conducted in work-related contexts, in which fairness has an objective meaning, e.g., treating people equally and making a diligent effort when contributing to shared activities [7]. In the context of teamwork, agents were found to be rewarded for being cooperative and fair and punished for being

selfish, and the perception of an agent's fairness to be dependent on the agent's status and trustworthiness [5]. Analysis of fairness in resource allocation did not reveal significant differences in the assessment of human and agents' (un)fairness [6].

Studies have shown that the experience of agents and their behavior is strongly affected by anthropomorphism, and anthropomorphic qualities generally increase agents' likability and trustworthiness in a social context [8]. While anthropomorphism facilitates trust, it may, however, also result in more disappointment when agents do not meet human expectations [9]. It was also found that children are particularly positive toward certain agents' embodiments, such as child-like or dog-like robots [10], people tend to comply more with feminine intelligent agents rather than with masculine ones [3], and that there is a preference toward intelligent agents with anthropomorphic qualities and personalities matching persons' own cultures [11]. An increased human-likeness of an agent tends to positively correlate with the tendency to anthropomorphize [12] but the dependency is not linear: according to the "uncanny valley" hypothesis, too much human-likeness can be experienced negatively [13].

While existing research provides valuable insights regarding perceptions of agents' fairness, little is still known about these phenomena in non-work contexts (which are becoming increasingly central due to recent technological developments). In addition, the relation between fairness and anthropomorphism has been relatively unexplored.

To deal with these limitations, this study addressed the questions of *How is fairness experienced as an attribute of a nonhuman agent?* and *How is the experience of fairness affected by the agent's perceived level of anthropomorphism?* Accordingly, the study focused on game-like interaction, rather than interaction in a work-related context, and employed agents with the appearance of either humanoid robots or non-humanlike objects. The study comprised an experiment, implemented as a quiz game hosted by an agent, which was followed by a focus group, where the participants shared their thoughts and reflections regarding intelligent agents' fairness, grounded in their prior first-hand quiz game experience.

2 Study 1: Quiz Game Experiment

2.1 Method

Eight persons (4 females and 4 males), 18–35 y. o., with the median age group being 21–25 y. o., took part in the experiment.

The study used a two-factor within-subject design. A combination of two independent variables, "Treatment" (Fair/Unfair) of the participants and agents' "Anthropomorphism" (Low/High), produced four experimental conditions. Each participant was exposed to all four conditions, in an order determined by using the Latin Square technique.

Four agent identities corresponded to four conditions of the study. In "Unfair"—as opposed to "Fair"—conditions the agents asked the participants more difficult questions and sometimes did not reward a participant for being faster than the other player. Two "Low anthropomorphism" agents, named "XA-Q4" (the "Fair" condition) and "ZW-X3" (the "Unfair" condition), had the appearance of a regular speaker (Fig. 1a). Two "High Anthropomorphism" agents, "Sam" (the "Fair" condition) and "Kim" (the "Unfair"

condition), looked like a humanoid robot having a “body”, “head”, rudimentary “arms”, and “tracked legs” (Fig. 1b). An iPad serving as the robot’s “face” displayed light-blue eyes, blinking 15–20 times per minute, and a static mouth with a neutral expression. Google Cloud Text-to-speech was used to generate two human-like voices for Sam and Kim and two mechanical voices for XA-Q4 and ZW-X3. The assignment of voices within the pairs was switched halfway through the experiment.

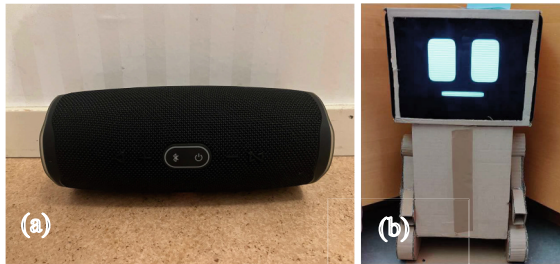


Fig. 1. (a) non-humanlike agents, XA-Q4 and ZW-X3, (b) humanoid agents, Sam and Kim.

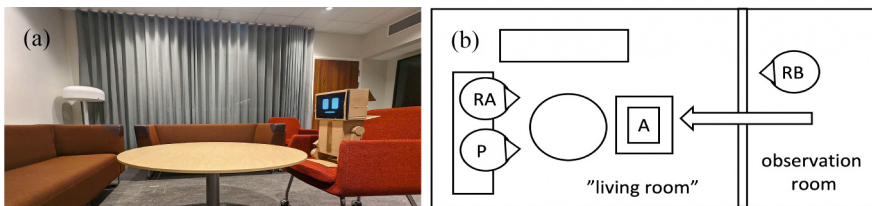


Fig. 2. The experimental setting: (a) a “living room” view, (b) the lab space floor plan (P: participant, A: agent, RA: Researcher A, RB: Researcher B)

The experiment took place at a lab space comprising a model “living room” and an adjacent observation room, separated by a one-way mirror (Fig. 2)¹.

The participants were informed about the aim of the study (which was described as assessing different types of socially aware intelligent agents) and provided their informed consent. Each participant then took part in four experimental conditions, acting as one of two players competing with one another in a quiz game “hosted” by an agent. The second player was enacted by Researcher A. The agents were controlled, using a Wizard of Oz (WoZ) technique, by Researcher B located in the observation room. Via its computer, Researcher B controlled a Bluetooth connected speaker, which was placed either on a table in front of the players or in the body of the agent. Both researchers followed a strict script designed to offer as little interactions outside of the script as possible. After each condition, the participants filled in a post-condition questionnaire. Each experimental session concluded with a semi-structured interview and debriefing. The sessions were about 60 min long.

¹ The first two authors conducted the experiment and focus group within their thesis work [14].

The post-condition questionnaire comprised ten attributes selected from Godspeed [15] and RoSAS [16] instruments, namely, *Happy/Feeling/Compassionate/Capable/Interactive/Reliable/Knowledgeable/Awkward/Awful/Social*, and three additional attributes, *Fair/Polite/Anthropomorphic*. All attributes were presented as 7-point Likert items. In post-interviews, the participants were asked about their experience in the study and their perception of robots' appearance and behavior. They were not explicitly prompted to talk about agents' fairness.

The study followed established guidelines for ethical research, including informed consent and confidentiality. The participants were debriefed about the use of the WoZ approach in the study.

2.2 Results

Post-condition Questionnaire. The results of post-condition questionnaire scores were analyzed using two-factor ANOVA. The "Fair" scores (Fig. 3a) showed a highly significant main effect of the "Treatment" factor ($F = 32,0627$, $P\text{-value} = 4,56E-06$, $F\text{ crit} = 4,196$), and the "Anthropomorphic" scores (Fig. 3b) showed a significant main effect of the "Anthropomorphism" factor ($F = 13,5318$, $P\text{-value} = 0,001$, $F\text{ crit} = 4,196$). The results served as a manipulation check: they confirm that independent variables were manipulated successfully, and differences between, respectively, "Fair vs. Unfair" and "High vs. Low Anthropomorphism" conditions were experienced by the participants as intended.

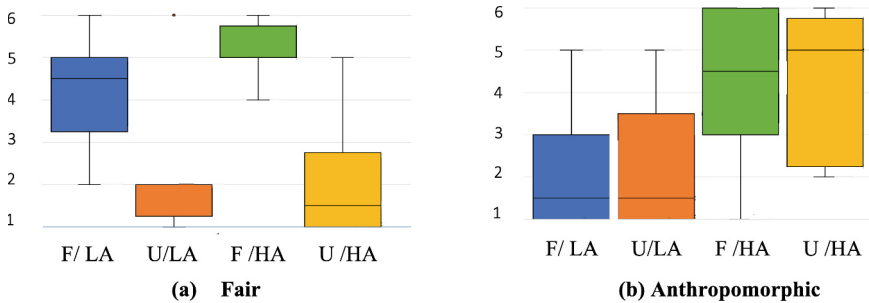


Fig. 3. "Fair" scores (a) and "Anthropomorphic" scores (b) in the conditions of the study (F: Fair, U: Unfair, LA: Low Anthropomorphism, HA: High Anthropomorphism).

There were no statistically significant interaction effects between the independent variables, and nor did we find significant main effects of the independent variables on the remaining scores. In particular, there were no significant differences between the conditions in the "Polite" scores and no significant differences between high- and low-anthropomorphism conditions in the "Social" scores.

Post-interviews. While the participants were generally positive about the experiment, finding it fun and interesting, they reported the feeling of being treated unequally. That could cause frustration, particularly because they could not appeal to an agent in the

same way they would do to a human. It was emphasized that agents should treat people equally and without bias, especially in competitive or collaborative contexts.

Somewhat surprisingly, explicit references to “fairness” were rather infrequent and general. At the same time, many detailed comments were provided on the agents’ appearance and behavior. The participants noted that while robots’ responsiveness and social awareness (e.g., greetings, playful comments) were limited, they still positively contributed to the feeling of social engagement. The participants also stated their general preference for expressive and anthropomorphic, but distinctly non-human, agents, and noted that agent’s voice is important for optimal user experience.

3 Study 2: Focus Group

3.1 Method

Five participants from Study 1 (4 females, 1 male) took part in Study 2.

The study was arranged as a face-to-face focus group session, explicitly focusing on the perception of technological agents’ fairness. The session involved two researchers and five participants. During the session, the participants were asked to: (a) reflect on their experience during Study 1 (in particular, whether they were treated unequally and/or unfairly), (b) share their personal views on fairness in general, (c) comment on three fictional “unfair agent” scenarios, illustrated by storyboards: a quiz host agent being unfair to one of the players (like in Study 1), household chores being unfairly assigned by an agent to family members, and an agent unfairly taking the side of its owner in a dispute with another person, and (d) present their opinion on the social rules that should be followed by intelligent agents. The session was about 60 min long.

The study followed established ethical research guidelines, including informed consent and confidentiality.

3.2 Results

The focus group discussion was recorded, transcribed, and analyzed using the thematic analysis framework [17]. The analysis produced the following main themes: *Perceived Fairness of Agents*, *Agents vs. Humans*, *Fairness vs. Politeness*, *Centrality of Context*, and *Fairness as a Design Objective*.

Perceived Fairness of Agents. The participants commonly described their experience in the experimental conditions in which they were treated unequally as “*it was unfair*”. They also occasionally mentioned *robots “being unfair”*. At the same time, they were hesitant to unreservedly describe the robots’ (and, in general, intelligent agents’) behavior in terms of “fairness”. The participants mentioned alternative explanations, namely, that (a) robots’ behavior was determined by how they were coded and trained rather than by the robots themselves, and (b) since *being fair* involves sympathy and empathy, and the robots lacked these qualities, they could not be considered “fair” (and, by extension, “unfair”). It was also mentioned that the experience of unfairness in the experiment could be more pronounced if “the stakes were higher”.

Agents vs. Humans. A recurrent topic of the discussion was a comparison of robotic and human game hosts. The participants observed that if a human host is unfair, one can confront them, but it is not clear how, or if, one can do that when the host is a robot. It was also noted that the robotic hosts in the quiz game experiment lacked the rich nonverbal cues, which are spontaneously produced by humans (e.g., “*A human would raise an eyebrow or show something on the face*”), and it made quiz-related interactions less dynamic.

Fairness vs. Politeness. According to the participants, the relationship between fairness and another key social attribute, politeness, is not straightforward. On the one hand, fairness and politeness are two different attributes, as, for instance, “*you can be rude and fair at the same time*”. On the other hand, while different, these attributes are not completely independent from one another. In particular, if someone is unfair, their behavior may be perceived as “disingenuous” or even “sarcastic”, rather than “polite”.

Centrality of Context. The participants emphasized that the perception of fairness, and, more broadly, social acceptability, of agents' behavior strongly depends on the context of interaction and the agent's role in the context. The same behavior may be acceptable in one context and unacceptable in another. For instance, if an agent approaches a person with a task assignment or a reminder, it can be natural aspect of activities in a work setting, but may not be perceived well outside work, especially if it happens in front of other people. And even within the same context, what is fair and acceptable may depend on the role of the agent: e.g., a scenario in which an agent spontaneously intervened in a human-human dispute was considered by the participants particularly unacceptable because the agent in the scenario took the side of its owner.

Fairness as a Design Objective. A common position, emerging from the focus group discussion, was that fairness and equality should be important objectives when designing intelligent agents and their behavior. There were, however, also concerns about how feasible this objective is, given that what is fair is subjective and context-specific, and therefore hard to define. Main suggestions regarding how to design intelligent agents to support their perceived fairness, proposed by the participants, can be summarized as follows. First, consistent with the experiment findings, participants favored agents that were expressive and capable of rich and dynamic interaction, but also distinctly non-human. Second, agents should be able to recognize the social context of their use and adjust their behavior to the context. Third, the design of agents should be based on a transparent set of rules and explicitly support human values, such as equality.

4 Discussion

The dual nature of technological agents as, on the one hand, social actors and, on the other hand, inanimate “things”, is one of the most intriguing and challenging issues in current social robotics research [18–20]. Evidence from our study highlights this duality. While the participants' rating scores and some verbal responses explicitly characterized the agents used in the study as “unfair”, it also transpired that the participants were skeptical about attributing unfairness to robots' autonomous agency. Concerns about

potential misunderstandings of robots as “true” agents appear to be a key reason behind the suggestion to design robots so that they are expressive but distinctly non-human.

The study highlights a particular aspect of the actor/ thing dilemma in human-agent interaction, namely, *limited possibilities for establishing a shared interpretation of the applicable social rules*. A crucial part of establishing a coherent social order in a setting is collaboratively deciding what is/ is not acceptable. If some actors disagree, they may appeal, complain, or even take the other party to court. With technological agents, it may not be possible. In our study, agents’ limited communication capabilities, somewhat paradoxically, were perceived as giving them more power over people. Since the participants were effectively prevented from conveying their complaints and requests to the agents, their overall experience of unfairness included not only being treated differently than the other player but also a lack of power to influence the agent.

The finding suggests that a potential direction for the design of human-agent interaction is making it possible for humans to collaboratively establish ‘house rules’ with agents. This design direction, which can capitalize on recent developments in LLMs and generative AI, should, however, be explored with caution to avoid placing humans in an unfavorable negotiation position and reinforcing unfairness in subtler ways.

In addition, while quantitative data from the quiz game experiment were mostly intended as a manipulation check, they tentatively suggest that agents’ “fairness” is a separate attribute, independent from both “anthropomorphism” (since we did not find a significant interaction between these factors) and “politeness” (since we did not find a significant effect of “Fairness” on the “Polite” scores). These hypotheses need to be verified in further studies.

5 Conclusion

The results of the exploratory study reported in this paper indicate that while fairness is experienced as an important aspect of human-agent interaction, operationalizing the notion of “agents’ fairness” is complicated because of its ambiguous and contextual nature and entanglement with other aspects of interaction. The evidence reported in this paper may inform the choice of methodology and help shape specific, testable research questions for future studies.

6 Disclosure of Interests.

The author has no competing interests to declare that are relevant to the content of this article.

Acknowledgments. The authors would like to thank study participants for their time and insights. This study was funded by The Swedish Research Council (grant number 2021-05409).

References

1. Sevillano, E.: Custom views and reception robots: This is what hotels of the future will look like. *El País*, April 20, 2025. <https://english.elpais.com/travel/2025-04-20/custom-views-and-reception-robots-this-is-what-hotels-of-the-future-will-look-like.html>
2. Dautenhahn, K.: Socially intelligent robots: dimensions of human-robot interaction. *Phil. Trans. R Soc. B Lond. B Biol. Sci.* **362**(1480), 679–704 (2007)
3. Siegel, M., Breazeal, C., Norton, M.I.: Persuasive robotics: the influence of robot gender on human behavior. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2563–2568. IEEE (2009)
4. Ötting, S., Gopinathan, S., Maier, G, et al.: Why criteria of decision fairness should be considered in robot design. In: Presented at CSCW 2017 (2017). <https://sites.coecis.cornell.edu/hri/files/2017/01/%C3%96tting-Gopinathan-Maier-and-Steil-2e7dua8.pdf>
5. Cao, J., Chen, N.: The influence of robots' fairness on humans' reward-punishment behaviors and trust in human-robot cooperative teams. *Hum. Factors* **66**(4), 1103–1117 (2024)
6. Claire, H., Kim, S., Kizilcec, R.F., et al.: The social consequences of machine allocation behavior: fairness, interpersonal perceptions and performance. *Comput. Hum. Behav.* **146**(C) (2023)
7. Chang, M.L., Pope, Z., Short, E.S., et al.: Defining fairness in human-robot teams. In: Proceedings of RO-MAN 2020, pp. 1251–1258. IEEE (2020)
8. Roesler, E., Manzey, D., Onnasch, L.: Embodiment matters in social HRI research: effectiveness of anthropomorphism on subjective and objective outcomes. *THRI* **12**(1), 1–9 (2023)
9. Waytz, A., Cacioppo, J., Epley, N.: Who sees human?: the stability and importance of individual differences in anthropomorphism. *Perspect. Psychol. Sci.* **5**(3), 219–232 (2010)
10. Syrdal, D.S., Koay, K.L., Walters, M.L., et al.: The boy-robot should bark! – Children's impressions of agent migration into diverse embodiments. In: Proceedings New Frontiers of Human-Robot Interaction, a Symposium at AISB (2009)
11. Korn, O., Akalin, N., Gouveia, R.: Understanding cultural preferences for social robots: a study in German and Arab communities. *THRI* **10**(2), 1–19 (2021)
12. Rothstein, N., Kounios, J., Ayaz, H., De Visser, E.J.: Assessment of Human-Likeness and Anthropomorphism of Robots: A Literature Review. In: *Advances in Neuroergonomics and Cognitive Engineering*. pp. 190–196. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-51041-1_26
13. Mori, M., MacDorman, K., Kageki, N.: The uncanny valley [from the field]. *IEEE Robot. Automat. Mag.* **19**, 98–100 (2012)
14. Bäckström, A., Ekenberg, W.: Don't be unfair, Mr Bot! An empirical study exploring the perception of fairness in non-work settings for human-agent interactions. Master's thesis. Department of Informatics, Umeå University (2023)
15. Bartneck, C., Kulić, D., Croft, E., et al.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **1**, 71–81 (2009)
16. Carpinella, C.M., Wyman, A.B., Perez, M.A., et al.: The robotic social attributes scale (RoSAS): development and validation. In: Proceedings of HRI 2017, pp. 254–262. ACM (2017)
17. Braun, V., Clarke, V.: Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**, 77–101 (2006)
18. Clark, H.H., Fischer, K.: Social robots as depictions of social agents. *Behav. Brain Sci.* **46**, e21 (2023). <https://doi.org/10.1017/S0140525X22000668>

19. Ziemke, T.: Understanding social robots: attribution of intentional agency to artificial and biological bodies. *Artif. Life* **29**, 351–366 (2023). https://doi.org/10.1162/artl_a_00404
20. Kaptelinin, V., Dalli, K.C.: Understanding contextual framing: a nonessentialist perspective on social interactions with technological artifacts. In: Proceedings of HRI 2025, pp. 1121–1130. IEEE, (2025). <https://doi.org/10.1109/HRI61500.2025.10974062>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

