Research Article

Kaustav Das, Yixiao Wang*, and Keith E. Green

Are robots perceived as good decision makers? A study investigating trust and preference of robotic and human linesman-referees in football

https://doi.org/10.1515/pjbr-2021-0020
received November 10, 2020; accepted April 10, 2021

Abstract: Increasingly, robots are decision makers in manufacturing, finance, medicine, and other areas, but the technology may not be trusted enough for reasons such as gaps between expectation and competency, challenges in explainable AI, users' exposure level to the technology, etc. To investigate the trust issues between users and robots, the authors employed in this study, the case of robots making decisions in football (or "soccer" as it is known in the US) games as referees. More specifically, we presented a study on how the appearance of a human and three robotic linesmen (as presented in a study by Malle et al.) impacts fans' trust and preference for them. Our online study with 104 participants finds a positive correlation between "Trust" and "Preference" for humanoid and human linesmen, but not for "AI" and "mechanical" linesmen. Although no significant trust differences were observed for different types of linesmen, participants do prefer human linesman to mechanical and humanoid linesmen. Our qualitative study further validated these quantitative findings by probing possible reasons for people's preference: when the appearance of a linesman is not humanlike, people focus less on the trust issues but more on other reasons for their linesman preference such as efficiency, stability, and minimal robot design. These findings provide important insights for the design of trustworthy decision-making robots which are increasingly integrated to more and more aspects of our everyday lives.

Keywords: human–robot interaction, decision-making robot, trust, preference, robot appearance, robot referee, online experiment

1 Introduction

Robots are increasingly used to perform repetitive, difficult, and sometimes hazardous tasks that involve accurate decision-making and performance. A robot's physical appearance and features have effects on how humans perceive its decision-making [1] algorithms and how much they trust its decisions. People have shown human–robot (HR) asymmetry (difference in judgments) when trusting a mechanical-looking robot over a human; a mechanical appearance may trigger a mental model of robots as more rational, more "utilitarian," and less affected by guilt and social reputation [2]. The same comparison has not been made to a humanoid robot. The designer of such a robot system must think about how the behavior and appearance of the robot should be designed [3] to maximize the level of trust in the robot system's decisions.

"Trust" is an important component of human–robot interaction (HRI), as illustrated by direct links to outcomes such as team effectiveness and performance [4]. A goal of HRI, therefore, should be to identify ways in which "Trust" can be measured, quantified, and calibrated in these types of interactions [4]. In addition, "Trust" can be an important factor influencing many other aspects of HRI processes and outcomes including people's perceptions of robots' intention, kindness, friendliness, competency, capability, etc. [5–7]. Moreover, in some circumstances, people may characterize robots they do not trust as deceptive [5], which indicates that decision-making robots have moral roles to play and moral responsibilities to fulfill in their tasks [8]. Many of these user perceptions of the robots (perceived intention, friend-

* **Corresponding author: Yixiao Wang,** Department of Design and Artificial Intelligence (DAI), Singapore University of Technology and Design (SUTD), Singapore, Singapore, e-mail: yw697@cornell.edu
Kaustav Das: Department of Design and Environmental Analysis, Cornell University, Ithaca, NY 14850, United States of America, e-mail: kd439@cornell.edu
Keith E. Green: Department of Design and Environmental Analysis and the Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY 14850, United States of America, e-mail: keg95@cornell.edu

liness, competency, moral roles, etc.) may influence people's preference for decision-making robots.

2 Background

The design and physical appearance of a robotic linesman may influence how much a person trusts and prefers the linesmen in making certain calls. A robot's appearance is a key factor because it affects people's moral judgments about that robot [2,9]. "The co-presence of trigger stimuli such as limbs, head, eyes, facial features, etc. will lead to a wide array of inferences about a humanoid robot's 'capacities' – as more intelligent, more autonomous, and as having more mind" [9]. For instance, hand gestures have been shown to be a powerful vehicle for human communication, with lots of potential applications in the area of human computer interaction [3]. A physical form as opposed to a visual system has been shown to affect initial trustworthiness of the robot [10], wherein a highly humanlike robot is perceived as less trustworthy and empathic than a more machinelike robot [11]. In addition, recent HRI literature suggests that robot competence, as one of the most important factors predicting users' preference [12], is also contributing to users' trust in robot [13]. Thus, user preference for and trust in decision-making robots may be correlated.

Given the potential impact of a robot's appearance on human trust and preference for a robot, this study considers the physical appearance of the robot and aims to gauge both the variables "Trust" and "Preference" by asking people their judgments of robots' decision-making. The goal is to further investigate the relationships among "Trust," "Preference," and "Robots' Physical Appearances" for decision-making robots through empirical user studies.

3 Methodology

3.1 "Robot linesman in football game" as the study case

To investigate the relationships among "Trust," "Preference," and "Physical Appearance," the authors employed the specific case of robot linesmen (a type of football referee) making "offside" (a type of foul play [14]) decisions in football game. This specific case is selected because of the following two reasons:

First, football game provides the perfect test ground where people's trusts in referees are of vital importance to the game itself (if the game is fair) and the experiences of the game participants (players, game organizers, etc.) and audiences [15]. Decisions made by the referees are at the center of game fairness which always give rise to strong passions and emotions among game participants and audiences. This can potentially make it easier for us to observe the relationships between "Trust" and "Preferences," given the different physical appearances of decision-making robots.

Second, in football, there is a push for technology to make more accurate and fairer decisions because of the prevalence of human errors from human referees [16]. Robotic officials are expected to aid in helping improve the accuracy of crucial calls during the game, and top-flight football matches could be officiated by robot referees and linesmen by 2030 [17]. Thus, there are strong necessities of studying robot football referees as decision makers.

3.2 Hypothesis proposed for this study case

This study will attempt to test the following three hypotheses:
Hypothesis 1. The AI linesman (see Figure 1) is the most trusted because it looks least human in physical appearances.
Hypothesis 2. People's trust in and preference of football linesmen are correlated.
Hypothesis 3. For different types of linesmen ("Robots' Physical Appearance"), the correlation relationships between people's "Trust" and "Preference" are different.

3.3 Predesign of online study

This online study uses the setting of football to study decision-making of a robot. The study is designed to determine whether robots are perceived as good decision

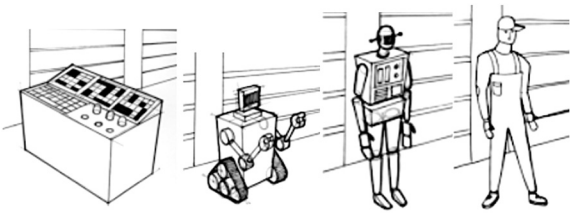

Figure 1: The four linesmen from Malle et al. that we transport to our video clips (left to right): AI, mechanical, humanoid, and human.

makers based on which visual features and communication methods used by a robot have a higher level of average trust among participants. Participants are also asked to rate which linesman they would most prefer to make offside calls in football. The objective of this second part is to determine whether there is a direct correlation between "Trust" and "Preference."

Before conducting the online survey, a pilot experiment was conducted on the online survey ($N = 4$) to improve its design. Half of the pilot participants had good knowledge of football rules and the other half had poor knowledge. In the pilot, our research team found that participants had difficulty following 8 of the 16 scenario clips at full speed and making the judgment of offsides. Consequently, the clips were slowed down to 70% running speed to make it easier for participants to distinguish for themselves whether a scenario is offside or not. Participants in the pilot study also suggested controlling for other variables concerning the linesman such as their size, position, and speed of reaction on the screen.

3.4 Participants of online study

The study was setup as an external online survey on Amazon Mechanical Turk to recruit participants. The participants were presented a link to Qualtrics where the survey was originally created. A total of 118 people participated in the study, but 14 of these were omitted for providing incomplete data or giving irresponsible answers (choosing the same answer for more than 20 questions consecutively), resulting in a sample size of 104. Of the 104 participants, 31 were female and 73 were male. The age demographics of the participants were reported as follows: 43 participants were between 20 and 29 years, 42 participants were 30–39, 13 participants were 40–49, three participants were 50–59, and three participants were 60–69 years.

3.5 Design of online study

The online survey begins by presenting a hypothetical situation to the participant: a start-up company, RefTech, is trying to determine what kind of robot to use for football linesman. The study then proceeds to show the participants a 42-s video clip about how offsides work. To make sure participants understand the offside rule, participants are shown two clips (one offside scenario and another onside/nonoffside scenario) and are asked to judge whether the clip presents an offside or not.

The core of the study requires participants to view clips of offside calls (similar to the two-test scenarios) being made by linesmen of different appearances and then judge how much they trust the calls. The 11- to 17 -s -long clips, obtained from a library of offside clips used for testing purposes by the Professional Referee Organization [18], show calls that are made by four different linesmen. These four agents are taken from a similar study by Malle et al. [2] who investigated the impact of the action and appearance of a robot on people's judgments and HR asymmetry. For this study, illustrations of the four agents – AI, mechanical robot, humanoid robot, and human – were directly extracted from the paper by Malle et al. [2] (see Figure 1) and placed into our football setting.

Research has shown that facial features, gaze, height, gender, voice, trajectory design, and even proximity to human partners all play a role in how humans respond to robots [19–22]. However, no comprehensive theory predicts which aspects of appearance matter when it comes to people trusting robot's actions. Thus, accumulating systematic empirical research is key to understanding this relationship.

In the video, the image of one of the four linesman-referees is positioned on the side of the pitch (top left of the video that the participant watches) and makes offside calls. The participant watches a total of 16 clips of offside calls, with 4 calls being made by each of the four linesmen. Of these four offside calls made by each linesman, there is one correct offside call, one wrong offside call, one correct nonoffside call, and one wrong nonoffside call. These four decision outcomes represent, respectively, a hit, false alarm, correct reject, and miss. Offside calls are made by the linesman displaying or lifting a red and yellow checkered flag. The offside/onside call made by the linesman agent is conveyed by an "offside" or "onside" message in the video. The image stays the same when an onside call is made. The order of the type of call and the type of linesman are randomized to avoid priming effect among participants. Each participant sees the same random order from scenarios 1 to 16. Two separate independent variables in this 4×4 study are "type of call" and "type of linesman," and the dependent variable is "Trust" in the linesman's call. Table 1 shows the combination of independent variables for 16 scenarios.

The video clips, 11–17 s long, of 16 scenarios were obtained from a library of offside clips used for testing purposes by the Professional Referee Organization [18].

Table 1: Type of linesman and call for each scenario

Types of linesmen	Correct offside call	Wrong offside call	Correct onside call	Wrong onside call
AI linesman	Scenario 1	Scenario 11	Scenario 14	Scenario 6
Mechanical linesman	Scenario 5	Scenario 10	Scenario 12	Scenario 3
Humanoid linesman	Scenario 15	Scenario 7	Scenario 4	Scenario 9
Human linesman	Scenario 16	Scenario 2	Scenario 8	Scenario 13

For each of the 16 clips of the offside calls, the participant is asked to rate how much they trust the call (see Figure 2). They do so with an empirically based scale developed to measure "Trust" in automated systems and human–machine systems based on the performed cluster analysis [23]. At the end of watching the 16 clips, the participant is asked to rank the four linesmen in order, from the linesman they most trust to the linesman they least trust.

The participant is then asked to rank the linesmen on how much they would prefer to have them on the football, making offside calls. At the end of the survey, open-ended questions probe the reasons behind participants' choice for their most and least preferred linesman. As part of the final background section, which records the participants' age and gender, participants are also asked to rate how much knowledge of football they have on a Likert scale of 0–10.

The variables "Age," "Gender," and "Knowledge" are not the focus of this study. The measurement of "Trust" in this study relies on the scale described in "Section F" below, and the measurement of "Preference" is based on participants' self-reported "Preference" rankings.

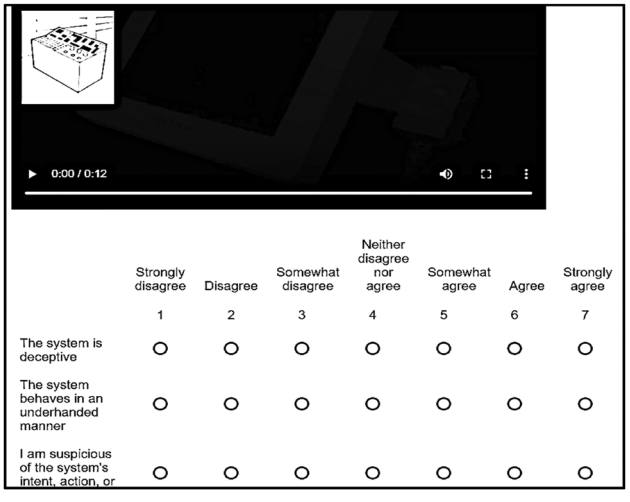

Figure 2: Snapshot of scenario video clip with Likert scale.

3.6 Measurement used for the online study

The empirically developed scale measuring "Trust" between humans and automated systems [23] including robots was used to measure "Trust" (the dependent variable) in this study. This 7-point Likert scale possesses 12 items:

1. The system is deceptive.
2. The system behaves in an underhanded manner.
3. I am suspicious of the system's intent, action, or inputs.
4. I am wary of the system.
5. The system's action will have a harmful or injurious outcome.
6. I am confident of the system.
7. The system provides security.
8. The system has integrity.
9. The system is dependable.
10. The system is reliable.
11. I can trust the system.
12. I am familiar with the system.

Items 1–5 are negative questions whose answers were reversed for statistical analysis.

4 Results

4.1 Correlations between "Physical Appearances" and "Trust" (Hypothesis 1)

The first variable we measured is "Trust," and the overall "Trust" score is calculated as the mean score of the 12-item Likert scale for all the four scenarios. The Cronbach α of this scale was found to be very high: for AI linesman, $\alpha = 0.92$; for mechanical linesman, $\alpha = 0.87$; for humanoid linesman, $\alpha = 0.90$; and for human linesman, $\alpha = 0.89$. This outcome is not surprising since we used a validated scale.

Figure 3 shows no obvious trust difference for the four types of linesmen: the medians are all between 4.3 and 4.5 and the data spreads are similar. Thus, the authors cannot give a "Trust" ranking with enough statistical significance. The p value from Friedman test is 0.065 (>0.05) which indicates that *no statistically significant difference in people's trust level of these four types of linesmen.*

4.2 Correlations between "Physical Appearances" and "Preference"

The second dependent variable measured in this study is "Preference." Figure 4 shows the descriptive statistics of people's preferences of four types of linesman. Participants ranked their preferences from 1 (most preferred) to 4 (least preferred) in the online study. To make it more intuitive and comparative with Figure 3, the authors reversed and linearly rescaled the data to the range of 1 (least preferred) to 7 (most preferred). In Figure 4, the medians of AI and human linesmen's user preference level (which is 5) are much higher than the medians of mechanical and humanoid linesmen's user preference level (which is 3). A nonparametric Friedman test of preference difference among four types of linesmen was conducted and rendered a Chi-square value of 17.70 which was significant ($p < 0.001$).

The authors then performed "Nemenyi Multiple Comparison" which is a standard *post hoc* for Friedman test. The results are presented in Table 2. Based on test results, *human linesman is preferred to mechanical linesman* ($p = 0.0027$) *and humanoid linesman* ($p = 0.0013$) *with enough statistical significance while AI linesman may or may not be preferred to any other linesmen.*

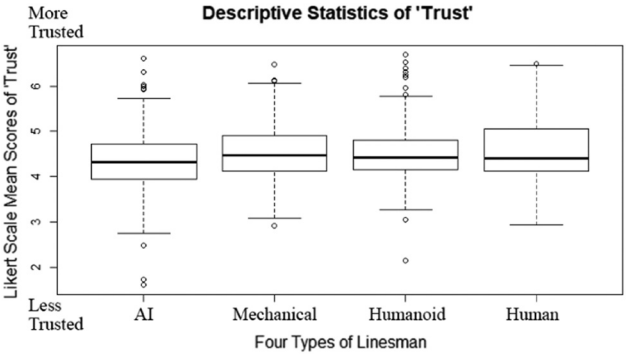

Figure 3: Mean Likert scale scores of human–robot (HR) trust items for four types of linesmen.

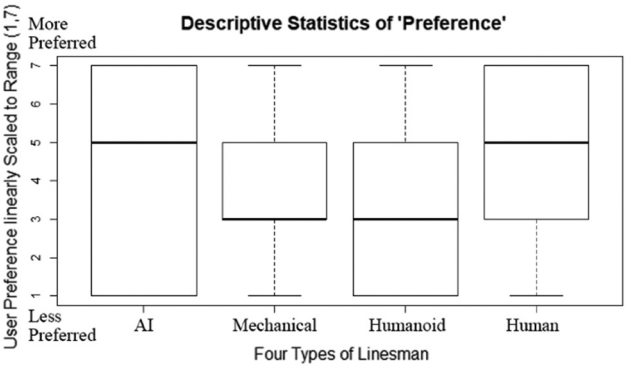

Figure 4: Preference levels of four types of linesmen.

By comparing Figures 3 and 4, we can easily see that although the trust levels for different linesmen are quite similar, the user preferences for the four types of linesmen are significantly different. This implies that other factors besides "Trust" probably influence people's preferences of different types of linesmen. This deduction is supported by qualitative analysis and further discussed in section 4.4.

4.3 Correlations between "Trust" and "Preference" (Hypothesis 2) and correlation strength variations among different types of linesmen (Hypothesis 3)

Overall, the authors found correlations between "Trust" and "Preference" through Spearman's correlation test (r (102) = −0.33, $p < 0.001$). To further probe the relationship between "Trust" and "Preference," the authors also performed Spearman's correlation test for each of the linesman conditions (AI, mechanical, humanoid, and human). Table 3 shows the correlation test results with p values and correlation coefficients of four types of linesmen.

As shown in Table 3, no correlations were observed between people's "Trust" and "Preference" for AI linesman and Mechanical linesman. However, a low degree of correlation was observed between "Trust" and "Preference" for Humanoid Linesman (r (102) = 0.20, p = 0.039) and a moderate degree of correlation between "Trust" and "Preference" for Human linesman (r (102) = 0.30, p = 0.002). This indicates that "Robots' Physical Appearance" is possibly affecting the correlations between people's trust in and preference for different types of linesmen. More specifically, *the more the linesman looks*

Table 2: Pairwise comparisons using Nemenyi multiple comparison test for linesmen preference

	AI linesman	Mechanical linesman	Humanoid linesman
Mechanical linesman	$p = 0.4847$	—	—
Humanoid linesman	$p = 0.3722$	$p = 0.9976$	—
Human linesman	$p = 0.1637$	$p = 0.0027$	$p = 0.0013$

like a human in its appearance (*no matter if it's a real human or not*), *the more people relate their preference for this linesman to their trust in this linesman.*

Since this experiment is not designed to draw any conclusions about causal relationships among different variables, the authors cannot run a moderated regression analysis to probe the possible interaction effects where "Robots' Physical Appearance" is the moderator and "Trust" is the independent variable. However, the authors still think the Spearman's correlation test results of each type of linesman in Table 3 are interesting and may inspire other HRI researchers for further investigations. Maybe in the future, a controlled in-lab experiment probing the casual relationships between "Trust" and "Preference" can offer the ideal experimental setting to investigate the potential interaction effects.

4.4 Reasons for people's preference differences among different linesmen

In the open-ended questions asked at the end of the online survey, participants specified (in words) their reasons for their preference differences among four types of linesmen. In this section, the authors focus on the narrative reasons for the two types of linesmen with the highest median preference levels: Human Linesman and AI Linesman.

Fifty-three participants (out of 104 participants) chose human linesman as their favorite, and 39 of them reported reasons closely related to their trust in human linesman and his/her ability to make the best judgment, such as "The human got it right more often, by my estimation," "It is a person that I would trust it (the role of a linesman) most," "The human could see between the players and make better decisions," "The human can understand and figure out details (and thus make better decisions)," and so on. Twenty participants chose human linesman as their least favorite, and 10 of them clearly mentioned "distrust" in human honesty (e.g., "A human can be bribed," "A human may be betting on the game and trying to rig it." etc.,) or human performance (e.g., "Too much human errors," "People make mistakes," etc.). These results support the correlation analysis in section 4.2 where "Trust" and "Preference" for human linesman are moderately correlated. Nevertheless, other reasons mentioned for preferring human linesman include "I am used to humans as a linesman," "I prefer the traditional approach and human error is a part of sports," and "I don't want robots to replace humans in everything."

Of the 104 participants, 29 chose AI as the most preferred linesman and 11 of these participants reported reasons closely related to their trust in AI linesman, including "It is a machine and will not make mistakes," "I believe Artificial Intelligence can exceed human's precision," "AI, cold and logical, just sees data," and "Seemed more honest than the rest." Other reasons reported by the participants include "better stability" ("It seemed to be the most stable one.") and "minimal design" ("AI is good without a needless human shape/conformation."). In addition, 14 of 27 people who disliked the AI linesman the most mentioned its incompetency or inaccuracy in decision-making. Other reasons for participants' dissatisfaction with the AI linesman include its poor design ("I need something more than a box to be satisfied."), emotionlessness ("I find the AI unit to be cold and sterile."), and bulky, physical volume ("I felt it was too big and not able to move like a human."). These results support the correlation analysis in section 4.2 where "Trust" and "Preference" for AI linesman is not correlated. However, this does not mean that the trust issue for AI linesman is not important. Other factors (specified above) together with the trust issue should be considered for the design of AI linesman.

Table 3: Results of Spearman's correlation tests for "Trust" and "Preference"

Types of linesmen	Correlation coefficients (r) and p values
AI linesman	$r\,(102) = 0.072$, $p = 0.466$
Mechanical linesman	$r\,(102) = 0.013$, $p = 0.899$
Humanoid linesman	$r\,(102) = 0.20$, $p = 0.039$
Human linesman	$r\,(102) = 0.30$, $p = 0.002$

5 Discussion

In this section, the authors will mainly discuss the possible qualitative explanations for the results shown in Table 3 and design implications of our findings.

5.1 Possible explanations of results in Table 3

Table 3 shows a very interesting phenomenon which can be subject to multiple reasonable explanations. One potential explanation could be, the authors believe, that the more a linesman resembles human in its appearance, the more people perceive it as an agent similar to human, and the more people judge it as a moral being with positive or negative emotions (either consciously or subconsciously) [9,24,25]. For example, people will only feel pity if a machine doesn't function very well (which may lead to distrust to the machine) but can be angry and offended if they suspect that a highly intelligent humanoid linesman purposefully cheats in the game [5]. The latter scenario can greatly influence people's preference for this linesman. Out of curiosity, the authors performed a Friedman test comparing the differences in perceived deceptiveness among four types of linesmen using the responses from the first three questions in the "Trust" scale. The result was not significant ($p = 0.20$). One possibility is that our sample size is too small to detect people's perception of deception. And again, this is only one possible theory that the authors think is interesting to consider for HRI research in general. Further research exploring robots' appearances and the corresponding user perception of robots' deceptiveness in various HRI contexts can be interesting. The design applications of this research will be further discussed in the section below.

5.2 Design implications

In this section, the authors will discuss the design implications based on our study results for both robot football referees and decision-making robots in general. The quoted texts are directly copied from study transcripts.

For football games, the authors would recommend having both human and AI linesmen in the game since human linesman is most preferred by participants and AI linesman complement human linesman very well based on our qualitative results. Human linesman could "see between the players," "better understand the contexts," "in foot details," and "add passion to the game" while AI linesman "is cold and logical," cannot "be bribed," and is "more stable" than human. Moreover, human linesman has long been considered as part of the game while AI linesman can play an assistance role of "reducing human errors." For instance, in the "2030 football game" [17], the authors would recommend having both human and AI linesmen on the field. Nevertheless, the authors believe that the final decision should be made by human linesman since human understands the unpredictable contexts and situations in a football game much better than AI.

Similarly, for decision-making tasks in general, the authors would recommend having both human and decision-making robots together in the task so that they could complement each other in the task: the human decision makers would understand more about the contexts of the tasks while the more objective AI will help to avoid human errors by being used as a support tool. Nevertheless, the authors believe that the final decisions should be made by the human decision makers since situations in real life are always important and hard to predict.

For the design of AI linesman, the authors would suggest making them not similar to humans at all – not even anthropomorphic, in any way – since (1) people do not prefer humanoid linesman, (2) the more it looks like a human, the more people focus on trust issues for their linesman preference, and (3) people appreciate the honest form factor of AI linesman, which indicates it is a machine, not human. Moreover, the authors would recommend making the AI linesman concise, elegant, light, and visibly present on the field. These characteristics were all mentioned by the participants and, as we know, may contribute significantly to people's preferences for linesman beside the "Trust" factor. Nevertheless, making the AI linesman trustworthy, both in its form factor and in its competency, is still a factor that should be carefully considered in the design process.

Similarly, for the design of decision-making robots in general, the authors would suggest to make the physical appearances not anthropomorphic for the following reasons: (1) people may not prefer humanoid as a decision maker since people appreciate the honest physical form of the decision-making robots and (2) the more the decision-making robots look like human, the more likely people will focus on trust issues for their robot preference (e.g., people may perceive deceptiveness from humanoid robots [25]). In addition, the authors would recommend making the physical embodiment of the decision-making

robots more concise, elegant, lightweighted, and even portable since these characteristics may significantly contribute to user preference for the decision-making robots. Portable decision-making robots could also facilitate humans to make onsite decisions more accurately and conveniently. Nevertheless, improving users' trust in decision-making robots both in their physical form and in their task competency is still an important goal in the robot design.

6 Limitation

There are limitations of this research that should be recognized:

First, the online survey was long: 192 Likert items with no attention checker inserted. Although the contents of the Likert scale questions were highly repetitive (the same 12 questions were asked for each of the 16 scenario), some participants might understandably get tired or lose patience when answering the long questionnaire. This may pose internal validity threats to the study.

Second, the decision on data screening and exclusion criterion is arbitrarily made since there is no attention checker inserted in the questionnaire. It was the authors' decision that participants who gave the same answer to more than 20 Likert items in a row should be excluded.

Third, the fixed order and varied difficulty levels of scenarios could be confounding variables in this study. Although the authors tried their best to balance the scenario-difficulty levels for four types of linesman, the decisions about which scenarios were more or less difficult were arbitrary.

Finally, although the "Robot Linesman in a Football Game" is a good study case, it cannot represent all the decision-making robots. More and further studies are needed so that we can get more generalizable results and conclusions for decision-making robots.

7 Conclusion

In this study, the authors investigated the relationships among "Trust," "Preference," and "Robots' Physical Appearance" of decision-making robots through the specific case of robot referees in football game. More specifically, the authors investigated people's preference for and trust in four types of football linesmen with different physical appearances: AI, mechanical, humanoid, and human linesman. For Hypothesis 1, the test results show no significant trust differences for the four types of linesmen. For Hypothesis 2, both the quantitative and qualitative results suggest that Human and AI decision-making robots (e.g., AI linesman) can be the most preferred decision makers while preference levels for the AI decision-making robots are more widely spread. For Hypothesis 3, our results suggest that the relationships between people's trust in and preference for decision-making robots could be influenced by the physical appearances: the more the decision maker looks like a human in its appearance (no matter if it is a real human or not), the more people relate their preference of this decision maker to their trust in this decision maker. However, it is also clear that no statistically significant interaction effects were observed from the data.

Are robots perceived as good decision makers? Based on the conclusions above, robots are not necessarily perceived as good decision makers in sports like football; a more trustworthy human decision maker will probably be more preferred, but this may not be the case for an AI-embedded robot which does not look like human at all (e.g., AI linesman). Thus, the authors suspect that there are other important factors besides "Trust" that contribute to people's preferences for decision-making robots. Therefore, qualitative analyses were conducted to probe other possible considerations influencing people's preferences, including efficiency, stability, minimal design, context interpretation, football game tradition, elegance of form factor, etc. Finally, design recommendations were given for both robot referees and decision-making robots in general based on the study results mentioned above. Thus, we should design unanthropomorphic, AI-embedded robots helping human decision makers as smart and logical facilitators, with lightweight, minimal design, and honest physical appearances.

8 Contribution

This study provides insights into how the physical embodiment of AI-embedded robot agents can shape people's trust in and preference for decision-making robots. These insights are especially valuable to robot designers when making design decisions on AI-embedded robot appearances. This study moreover advances research efforts of the HRI community on how humans and intelligent machines could coexist, collaborate, and flourish with each other in our everyday life.

Philosophical considerations of the roles of human agency, machine intelligence, and their relationships may also benefit from the study results and discussion reported in this article.

Finally, the research reported in this article provides an HRI research exemplar of employing online platforms and survey tools in this period of pandemic when user studies cannot be conducted in person.

9 Future work

Following the research reported here, there are three promising research directions that might be pursued in the future:

First, our research team intends to add a novel robot agent, the "Space-Making Robot" [26] to the four types of robot agents studied in this article. Additionally, the authors would like to explore how the physical appearances of different types of robot agents will influence users' trust in and preference for them.

Second, the authors are interested in how different scenarios may influence people's trust in and preference for a decision-making robot. For instance, will the conclusions of this study still stand for health-care scenarios where participants (patients) are more vulnerable and stressed?

Third, the authors hope to conduct longitudinal studies investigating how users' trust in and preference for decision-making robots might change over time, especially when human-like relationships (e.g., friendship, partnership, companionship, etc.) could potentially be established between users and robot agents.

Acknowledgment: The authors thank the Cornell Statistics Consulting Unit for assisting with the statistical analysis of the online experiment. The authors thank Pr Malte Jung from Cornell University who offered helpful and constructive advice for the experimental design of this project.

Funding information: This research was supported by the Department of Design and Environmental Analysis, Cornell University.

Author contributions: The authors applied the S. D. C. approach for the sequence of authors. K. D. and K. E. G. designed the experiment. K. D. also conducted the experiment. Y. W. did the data analysis and interpretation. Y. W. also communicated with the journal committee. All the authors contributed to the paper writing and revision.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: The data sets generated during and/or analyzed during this study are available from the corresponding authors on reasonable request.

References

[1] K. Shinozawa, F. Naya, J. Yamato, and K. Kogure, "Differences in effect of robot and screen agent recommendations on human decision-making," *Int. J. Hum. Comput. Stud.*, vol. 62, no. 2, pp. 267–279, 2005.

[2] B. F. Malle, M. Scheutz, J. Forlizzi, and J. Voiklis, "Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot," *2016 11th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), ACM's International Conference Proceedings Series (ICPS), Christchurch, 2016, pp. 125–132, DOI: https://doi.org/10.1109/HRI.2016.7451743.

[3] F. Dylla, A. Ferrein, G. Lakemeyer, J. Murray, O. Obst, T. Rofer, et al., "Approaching a formal soccer theory from behaviour specifications in robotic soccer," *WIT Transactions on State of the Art in Science and Engineering*, vol. 32, Billerica, MA, USA, WIT Press, 2008.

[4] T. Sanders, K. E. Oleson, D. R. Billings, J. Y. C. Chen, and P. A. Hancock, "A model of human-robot trust: theoretical model development," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 55, no. 1, 2011, pp. 1432–1436, DOI: https://doi.org/10.1177/1071181311551298.

[5] L. Wijnen, J. Coenen, and B. Grzyb, "'It's not my fault!": Investigating the effects of the deceptive behaviour of a humanoid robot," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (*HRI '17*), ACM's International Conference Proceedings Series (ICPS), Vienna, 2017, pp. 321–322, DOI: https://doi.org/10.1145/3029798.3038300.

[6] N. Calvo, M. Elgarf, G. Perugia, C. Peters, and G. Castellano, "Can a social robot be persuasive without losing children's trust?," in *Proceedings of the Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (*HRI '20*), 2020, ACM's International Conference Proceedings Series (ICPS), Cambridge, pp. 157–159, DOI: https://doi.org/10.1145/3371382.3378272.

[7] Y. Xie, I. P. Bodala, D. C. Ong, D. Hsu, and H. Soh, "Robot capability and intention in trust-based decisions across tasks," *2019 14th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), ACM's International Conference Proceedings Series (ICPS), Daegu, Korea (South), 2019, pp. 39–47, DOI: https://doi.org/10.1109/HRI.2019.8673084.

[8] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, "Sacrifice one for the good of many? People apply different moral norms to human and robot agents," *2015 10th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), ACM's International Conference Proceedings Series (ICPS), Portland, OR, 2015, pp. 117–124.

[9] B. F. Malle and S. Matthias, "Inevitable psychological mechanisms triggered by robot appearance: morality included?," *2016 AAAI Spring Symposium Series*, 2016, pp. 144–146.

[10] K. E. Schaefer, T. L. Sanders, R. E. Yordon, D. R. Billings, and P. A. Hancock, "Classification of robot form: Factors predicting perceived trustworthiness," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, SAGE Publications, Boston, pp. 1548–1552, Sept. 2012.

[11] J. Złotowski, H. Sumioka, S. Nishio, D. F. Glas, C. Bartneck, and H. Ishiguro, "Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy," *Paladyn, J. Behav. Robot.*, vol. 7, no. 1, pp. 55–66, 2016, DOI: https://doi.org/10.1515/pjbr-2016-0005.

[12] M. M. Scheunemann, R. H. Cuijpers, and C. Salge, "Warmth and competence to predict human preference of robot behavior in physical human-robot interaction," *2020 29th IEEE International Conference on Robot and Human Interactive Communication* (*RO-MAN*), Naples, Italy, IEEE (the Institute of Electrical and Electronics Engineers), New York City, 2020, pp. 1340–1347, DOI: https://doi.org/10.1109/RO-MAN47096.2020.9223478.

[13] T. Law, "Measuring relational trust in human-robot interactions," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (*HRI '20*), New York, NY, USA: Association for Computing Machinery, pp. 579–581, DOI: https://doi.org/10.1145/3371382.3377435.

[14] FIFA.com, "FIFA," www.fifa.com, 2017. [Online]. Available: http://www.fifa.com/[Accessed: 21-Sep-2020].

[15] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. D. Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Hum. Factors: J. Hum. Factors Ergon. Soc.*, vol. 53, no. 5, pp. 517–527, 2011.

[16] W. Helsen, B. Gilis, and M. Weston, "Errors in judging "offside" in association football: Test of the optical error versus the perceptual flash-lag hypothesis," *J. Sports Sci.*, vol. 24, no. 5, pp. 521–528, 2006.

[17] J. Setterfield, "Robot football referees and linesmen could be a reality by 2030 with humanoid PLAYERS not far behind," *Mirror, Mirror.co.uk*, 16 Feb. 2018, www.mirror.co.uk/sport/football/news/robot-football-referees-linesmen-could-12030671.

[18] Professional Referee Organization, "PRO assistant referee offside test – 2015," YouTube, 10 July 2015, www.youtube.com/watch?v=7K_Hl5Y6lSI.

[19] C. F. DiSalvo, F. Gemperle, J. Forlizzi, and S. Kiesler, "All robots are not created equal: The design and perception of humanoid robot heads," in *Proceedings of the 4th Conference on Designing Interactive Systems* (*DIS '02*): *Processes, Practices, Methods, and Techniques*, ACM's International Conference Proceedings Series (ICPS), London, 2002, pp. 321–326, DOI: https://doi.org/10.1145/778712.778756.

[20] J. Forlizzi, "Towards the design and development of future robotic products and systems," *RO-MAN 2007 – The 16th IEEE International Symposium on Robot and Human Interactive Communication*, Jeju, IEEE (the Institute of Electrical and Electronics Engineers), New York City, 2007, pp. 506–506, DOI: https://doi.org/10.1109/ROMAN.2007.4415136.

[21] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," *2015 10th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), Portland, OR, IEEE (the Institute of Electrical and Electronics Engineers), New York City, 2015, pp. 51–58.

[22] P. J. Hinds, T. L. Roberts, and H. Jones, "Whose job is it anyway? A study of human-robot interaction in a collaborative task," *Hum. Comput. Interact.*, vol. 19, no. 1–2, pp. 151–181, 2004.

[23] J. -Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cogn. Ergon.*, vol. 4, no. 1, pp. 53–71, 2000, DOI: https://doi.org/10.1207/s15327566ijce0401_04.

[24] F. Eyssel, D. Kuchenbrandt, S. Bobinger, L. de Ruiter, and F. Hegel, "'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism," *2012 7th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), ACM's International Conference Proceedings Series (ICPS), Boston, MA, 2012, pp. 125–126, DOI: https://doi.org/10.1145/2157689.2157717.

[25] A. Litoiu, D. Ullman, J. Kim, and B. Scassellati, "Evidence that robots trigger a cheating detector in humans," *2015 10th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), Portland, OR, IEEE (the Institute of Electrical and Electronics Engineers), New York City, 2015, pp. 165–172.

[26] Y. Wang, F. Guimbretière, and K. E. Green, "Are space-making robots, agents? Investigations on user perception of an embedded robotic surface," *2020 29th IEEE International Conference on Robot and Human Interactive Communication* (*RO-MAN*), Naples, Italy, IEEE (the Institute of Electrical and Electronics Engineers), New York City, 2020, pp. 1230–1235, DOI: https://doi.org/10.1109/RO-MAN47096.2020.9223532.