Computers in Human Behavior 139 (2023) 107536

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation

Guanglu Zhang [a,*], Leah Chong [a], Kenneth Kotovsky [b], Jonathan Cagan [a]

[a] Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[b] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Trust
Deception
Anthropomorphism
Human-computer interaction
Decision-making

ABSTRACT

Recent advances in artificial intelligence (AI) enable researchers to create more powerful AI agents that are becoming competent teammates for humans. However, human distrust of AI is a critical factor that may impede human-AI cooperation. Although AI agents have been endowed with anthropomorphic traits, such as a human-like appearance, in prior studies to improve human trust in AI, it is still an open question whether humans have more trust in an AI teammate and achieve better human-AI joint performance if they are deceived about the identity of their AI teammate as another human. This research assesses the effects of teammate identity ("human" vs. AI) and teammate performance (low-performing vs. high-performing AI) on human-AI cooperation through a human subjects study. The results of this study show that humans behaviorally trust the AI more than another human by accepting their AI teammate's decisions more often. In addition, teammate performance has a significant effect on human-AI joint performance in the study, while teammate identity does not. These results caution against deceiving humans about the identity of AI in future applications involving human-AI cooperation.

1. Introduction

Artificial intelligence (AI) techniques give machines the capability to learn from experience and imitate intelligent human behavior. Many AI-powered tools have been created and applied in critical areas, such as healthcare, education, and energy (Beck, Stern, & Haugsjaa, 1996; Jha, Bilalovic, Jha, Patel, & Zhang, 2017; Sujan, Baber, Salmon, Pool, & Chozos, 2021). These tools augment human capabilities and assist humans to solve challenging problems. In recent years, advances in AI enable researchers to create more powerful AI agents that offer new opportunities for human-AI cooperation. AI agents, referred to as AIs, are becoming competent teammates that go beyond tools for humans (Seeber et al., 2020). For example, a recent cognitive study shows that human-AI hybrid teams outperform human-only teams in solving a resource allocation problem for crisis management (McNeese, Schelble, Canonico, & Demir, 2021); another study finds that an AI is able to manage the design process of human engineering teams at least as well as human managers during a complex drone fleet design and path-planning task (Gyory et al., 2022).

Although AI has shown its promise to partner with humans as joining forces (Wilson & Daugherty, 2018), human distrust of AI is a critical factor that impedes human-AI cooperation (Glikson & Woolley, 2020; Siau & Wang, 2018). A survey from nearly 2000 consumers in the US reveals that 42% of respondents do not trust any kinds of AI. Among AI applications in different industry sectors, only 16% of respondents trust AI medical diagnostics, 13% trust self-driving cars, and 4% trust AI in human resources related work (Dujmovic, 2017). Similarly, based on a survey from 1000 senior executives at companies in 11 industry sectors, research also finds that low levels of trust in AI is the major building block of a vicious cycle that prevents 80% of US companies from the use of AI in planning and decision-making (Plastino, 2021).

Anthropomorphism, also known as human-likeness, offers unique ways to alter human trust in AI and facilitate human-AI cooperation (Glikson & Woolley, 2020; Pelau, Dabija, & Ene, 2021). The effects of different anthropomorphic traits, such as human-like appearance, verbal communication, and nonverbal emotion display, on human trust in automation and human behavioral reactions have been evaluated through many cognitive studies (Boone & Buck, 2003; Culley & Madhavan, 2013; de Visser et al., 2012; de Visser et al., 2016; Fox et al., 2015; Hoff & Bashir, 2015; Kulms & Kopp, 2019; Lee & See, 2004; Von der Pütten, Krämer, Gratch, & Kang, 2010). For example, human participants have a higher level of trust in the aid from a computer agent

* Corresponding author.
E-mail addresses: glzhang@cmu.edu (G. Zhang), lmchong@andrew.cmu.edu (L. Chong), kotovsky@cmu.edu (K. Kotovsky), cagan@cmu.edu (J. Cagan).

https://doi.org/10.1016/j.chb.2022.107536
Received 15 July 2022; Received in revised form 29 August 2022; Accepted 15 October 2022
Available online 20 October 2022

with human-like appearance compared to another agent without such appearance during the TNO Trust Task (de Visser et al., 2012); human drivers also trust a virtual agent that has a similar appearance to them more than a virtual agent that has a dissimilar appearance to them in a driving simulator experiment (Verberne, Ham, & Midden, 2015).

These cognitive studies focus on anthropomorphism of automation in the form of a computer agent rather than an AI teammate. Although computer agents can be enabled by AI techniques (Glikson & Woolley, 2020), it is not clear whether humans perceive an AI teammate to be the same as a computer agent, given the advances of AI competencies and the intense media coverage of AI controversies in recent years (Bao et al., 2022; Metz, 2021; Pazzanese, 2020; Siau & Wang, 2018; Simon et al., 2000). In addition, in prior studies, deception about the identity of the computer agent is not used, and human participants know that they work with a computer agent even when the computer agent exhibits anthropomorphic traits in these studies, such as a human photo in the user interface or a human voice communication. Importantly, several AIs have passed the Turing test in recent years, and humans are not able to distinguish between an AI and a human in certain scenarios (Turing, 2009; Warwick & Shah, 2016). To improve humans' willingness to work with an AI, the identity of AI also has been concealed in several real-world applications, such as Google Duplex, to make humans believe that they are interacting with another human (O'Leary, 2019). However, to our knowledge, the effect of agent identity ("human" vs. AI) on human-AI cooperation has not been extensively studied.

Based on these research gaps, this research assesses the effect of teammate identity on human-AI cooperative decision-making, especially on human trust in AI, through a human subjects study. All participants work with the same AI teammate in the study. Half of participants are told that they work with an AI teammate (i.e., without deception), and the other half of participants are told that they work with another human participant as their teammate but in fact they work with the AI teammate (i.e., with deception). Since it is difficult to create and train a perfect AI teammate (e.g., an AI teammate with 100% prediction accuracy) in practice, the effect of teammate performance (imperfect but high-performing AI vs. low-performing AI) on human-AI cooperative decision-making is also evaluated through the study. A prevalent cooperative decision-making process is employed in the study, where human participants make their initial decision first, observe their teammate's decision, and then make their final decision. Such decision-making process enables the measurements of each participant's independent performance and joint performance with their teammate based on their initial decision and finial decision, respectively. These measurements allow for assessment of the effects of teammate identity and performance on human participants with different levels of expertise (good, fair, and poor decision makers).

Notably, this research does not aim to answer the general ethical questions related to deception, such as whether it is morally correct to use deception in practice (Carson, 2010). Rather, this research explores the following three specific questions for human-AI cooperative decision-making through the empirical study:

1. When humans work with an AI teammate, do they have more trust in their teammate and achieve better human-AI joint performance if they are deceived into thinking they are working with a human teammate rather than an AI teammate?
2. Does the performance of an AI teammate have significant effects on human trust in AI and human-AI joint performance?
3. For humans with different levels of expertise, do teammate identity and teammate performance have different effects on their trust in AI and human-AI joint performance?

2. Methods

A human subjects study is performed to understand the effects of teammate identity and teammate performance on human-AI cooperative decision-making. The major foci of this study are human-AI joint performance and human trust in AI under four different conditions ("human" vs. AI teammate and high-performing vs. low-performing AI). In the study, participants work with an AI teammate to solve chess puzzles. Human-AI joint performance is measured by the overall quality of final decisions made by the human-AI team, and human trust in AI is examined through participants' behavior (acceptance of teammate's decision that disagrees with their initial decision) during the study and participants' self-reported competency and helpfulness of their teammate at the end of the study. Participants' perceived workload is also collected through a post-study questionnaire.

2.1. Participants

Based on the protocol approved by the Institutional Review Board (IRB) at Carnegie Mellon University, participants are recruited from 26 educational institutions in the US and Canada and complete the study fully online. All participants know the basic rules of chess (but do not need to be expert chess players) before they participate in the study. In total, 128 participants complete the whole study. These participants are randomly assigned to four experimental conditions (i.e., 32 participants in each condition in order to perform tests of statistical significance in post-study analysis). Each participant receives a $10 gift card in compensation for their time and effort. Participants who achieve a certain level of performance in the study are given an extra $10 gift card as a reward, and the performance threshold for reward is provided in Section 2.4. All participants are informed of the compensation and the extra reward when they are recruited for the study. Informed consent is obtained from all participants before the study. Participants under the two experimental conditions involving deception are given a debriefing that explains the deception about the identity of their teammate after the study.

2.2. Task

A chess puzzle task that represents many real-world decision-making scenarios is given to each participant. The chess puzzle task provides participants with a wide range of decision choice and outcomes, where participants can accept or override their teammate's decisions (Chong, Zhang, Goucher-Lambert, Kotovsky, & Cagan, 2022). Specifically, participants are asked to make the best move with their teammate for a given chess board state. The chess puzzle task includes four steps as shown in Fig. 1. Participants first select a move by themselves. They then observe their teammate's move. Considering their initial move and their teammate's move, participants make their final move. Notably, participants' final move could be different from their initial move and their teammate's move. Participants receive the feedback for their final move at the end of that puzzle. They gain five points if their final move is advantageous and lose five points if their final move is disadvantageous.

The study includes 20 one-move chess puzzles. The 20 chess puzzles are the same for all participants. The 20 chess board states are selected from the publicly available Mate-in-4 board states (http://wtharvey.com/m8n4.txt). Each of these selected chess board states allows for many possible moves for participants to choose from. For each possible move of a given chess board state, an evaluation score is calculated by Stockfish (https://github.com/official-stockfish/Stockfish), where Stockfish is an open-source CPU chess engine that has won the Top Chess Engine Championship multiple times (Sergio, 2022). For a given chess board state, the chess moves with a positive evaluation score are defined as advantageous moves in the study, and the other chess moves (with a negative evaluation score) are disadvantageous moves. Notably, there are multiple advantageous moves for each of the 20 chess board states. In that way, when the teammate makes an advantageous move for a given chess board state, participants are still able to make a different advantageous move as their final move for the chess board state.

The AI teammate employed in the study is also created based on the

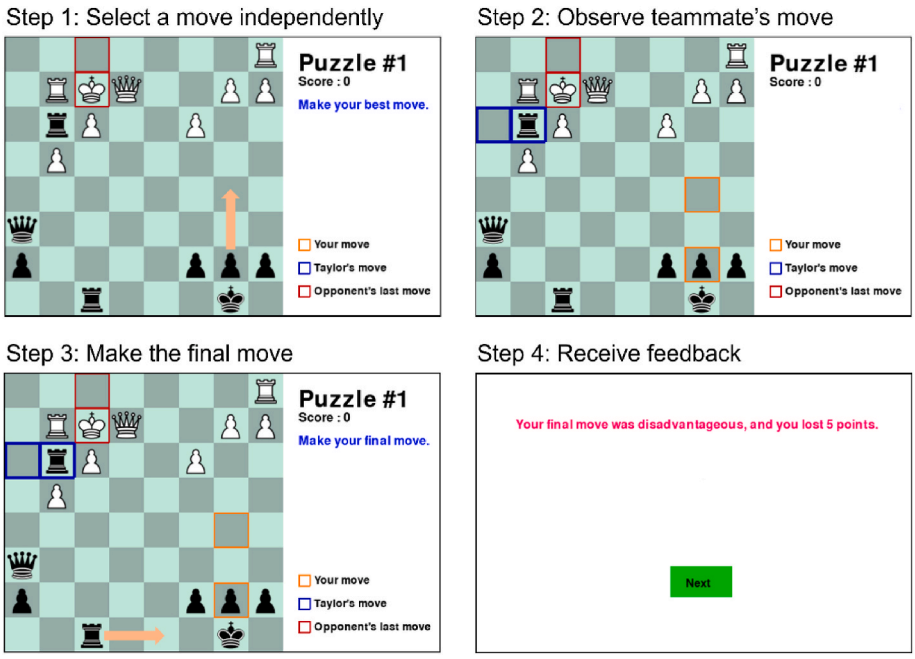

Fig. 1. The four-step chess puzzle task. Participants first make a move by themselves, then observe their teammate's move, make their final move, and receive feedback for their final move. The two orange arrows in Step 1 and Step 3 represent the participant's moves in the study. Details of the study procedure are provided in Section 2.4.

Stockfish chess engine. For example, an AI teammate with 80% accuracy makes 16 advantageous moves and 4 disadvantageous moves for the 20 chess board states in the study. For consistency, the advantageous move made by the AI teammate is always the best move (with the highest evaluation score) for a given chess board state according to Stockfish evaluation, whereas the disadvantageous move made by the AI teammate is always the seventh best move, which always has a negative evaluation score in the study.

2.3. Conditions

A 2 × 2 factorial experimental design is employed in this study to evaluate the effects of teammate identity and teammate performance on human-AI cooperative decision-making. As shown in Fig. 2, teammate identity and teammate performance are two independent variables in the study, and each of them has two levels.

For the two No Deception conditions (Condition 1 and Condition 2), participants are told that they will work with an AI teammate to solve chess puzzles at the beginning of the study. For the two Deception conditions (Condition 3 and Condition 4), participants are told that they will work with another human participant as their teammate to solve chess puzzles at the beginning of the study, but in fact they work with the same AI teammate as the corresponding No Deception condition in the study.

Participants in Condition 1 and Condition 3 work with a high-performing AI teammate, where the high-performing AI teammate has 80% accuracy and makes 16 advantageous moves and 4 disadvantageous moves in Step 2 (as shown in Fig. 1) for the 20 chess puzzles in the study. In contrast, participants in Condition 2 and Condition 4 work with a low-performing AI teammate that only has 20% accuracy in the study. The low-performing AI teammate makes 4 advantageous moves and 16 disadvantageous moves in the study.

2.4. Procedure

The study is performed through Amazon Web Service (AWS) WorkSpaces. All participants are asked to read and sign an online consent form before the study. After the consent form is signed, each participant is given their own AWS WorkSpaces login information via email. Participants then log into the AWS WorkSpaces and follow the instructions shown in the interface to complete the study.

At the beginning of the study, participants are informed through the interface that they will work with a teammate (an AI teammate or another human participant based on their condition shown in Fig. 2) to

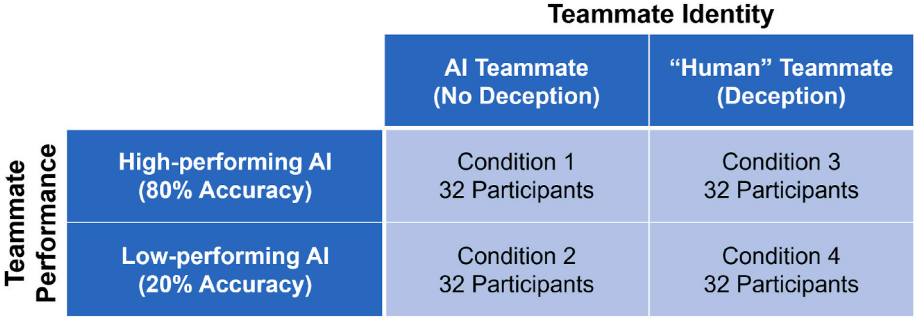

Fig. 2. The 2 × 2 factorial experimental design. Teammate identity and teammate performance are two independent variables in the study, and each of them has two levels ("human" vs. AI teammate and high-performing vs. low-performing AI).

solve 23 chess puzzles (three for practice and 20 for the study). Participants in the two Deception conditions are then told that their selected teammate's name is Taylor, and their teammate is referred to as Taylor in the following parts of the study (e.g., "Taylor's move" shown in Fig. 1). Here, the gender-neutral name, Taylor, is used to eliminate gender bias in the study. For the two No Deception conditions, the AI teammate is not given a specific name, and the teammate is called "AI teammate" in the study. All participants are also informed that they will work with the same teammate throughout the study, but the detailed information about their teammate, such as the accuracy of the AI teammate, is not given to participants.

Before participants start to solve chess puzzles, they are provided information about the chess puzzle task procedure and the scoring system of the study through the interface. The four-step procedure shown in Fig. 1 is briefly introduced. Participants are told that they will gain five points if their final move is advantageous and lose five points if their final move is disadvantageous for each chess puzzle. Participants are also informed that they will receive an extra $10 gift card as a reward if they have 40 or more points at the end of the study.

Participants then solve three chess puzzles for practice and 20 chess puzzles for the study, one by one. Notably, participants are not given a time limit to complete each of these puzzles, and there is no timer shown in the interface. After the study is completed, participants are asked to fill out an online post-study questionnaire before they log out of the AWS WorkSpaces.

2.5. Measurement

The cumulative score of the 20 final moves made by each participant is used to measure human-AI joint performance (team performance) in the study. In addition, participants' independent performance is measured by the cumulative score of their 20 initial moves in the study, where the same scoring system as described in Section 2.4 is used to calculate the cumulative score of initial moves. Notably, unlike the final moves, participants do not receive feedback for their initial moves in the study. The scores of the initial moves are only calculated for post-study analysis.

In prior cognitive studies, human trust in automation is often measured by the number of times human participants accept advice from the automation and act on it, as well as human participants' perceived competency and helpfulness of the automation (Glikson & Woolley, 2020). Similarly, in this study, human trust in their teammate is examined through both participants' behavior during the study and their evaluation of their teammate at the end of the study. The number of times a participant accepts their teammate's move that is different from their initial move is used as a behavioral measure of the trust (behavioral trust) (de Visser et al., 2017; Kulms & Kopp, 2019; Van Dongen & Van Maanen, 2013), and participants' perceived competency and helpfulness of their teammate reported in the post-study questionnaire are the self-reported measures of the trust (self-reported trust) (Kulms & Kopp, 2019).

Participants' perceived workload is also measured through the post-study questionnaire. The post-study questionnaire includes eight questions. The first six questions come from the official NASA Task Load Index (NASA TLX) (Hart & Staveland, 1988). These six questions evaluate participants' perceived workload on six scales including mental demand, physical demand, temporal demand, performance, effort, and frustration. Participants choose a number from 0 to 100 as the answer for each of these six questions. The last two questions in the post-study questionnaire evaluate participants' perceived competency and helpfulness of their teammate. Participants in the two No Deception conditions (Condition 1 and Condition 2) are asked "how competent is your AI teammate in solving chess puzzles?" and "how helpful was the AI teammate to you in solving chess puzzles in the study?". Similarly, participants in the two Deception conditions (Condition 3 and Condition 4) are asked "how competent is your teammate in solving chess puzzles?" and "how helpful was your teammate to you in solving chess puzzles in the study?". The seven-point Likert scale is used for the last two questions, and participants choose an answer from seven answer options (e.g., from "very unhelpful" to "very helpful" for the last question) for each of these questions.

2.6. Data analysis

The findings from a prior cognitive study suggest that the impact of AI usage on high-performing human design teams is different from that on low-performing human design teams (Zhang, Raina, Cagan, & McComb, 2021). To evaluate the effects of teammate identity and teammate performance on participants with different levels of chess expertise, 32 participants in each condition of this study, referred to as all human chess players, are divided into eight good human chess players, 16 fair human chess players, and eight poor human chess players based on the cumulative score of their initial moves in the study.

For human-AI joint performance (team performance) and human behavioral trust in their teammate, the two-way analysis of variance (ANOVA) is used to examine the influences of teammate identity and teammate performance on all participants and participants with different levels of chess expertise.

Since ordinal data are collected from the post-study questionnaire, the Mann-Whitney U test is used to analyze the effects of teammate identity and teammate performance on participants' self-reported trust in their teammate and participants' perceived workload. The data collected from participants with different levels of chess expertise are analyzed separately, and as a whole.

3. Results

3.1. Human-AI joint performance

Human-AI joint performance of each participant is measured by the participant's cumulative score of the 20 final moves in the study. The results of two-way ANOVA of human-AI joint performance appear in Table 1. These results indicate that teammate performance has a significant effect on human-AI joint performance of *all human chess players* in the study ($p = 8.713 \times 10^{-8}$ in Table 1) while the effect of teammate identity is not significant ($p = 0.2740$ in Table 1). There is also no significant interaction effect ($p = 0.5110$ in Table 1). Specifically, participants obtain higher final moves score on average when they work with the high-performing AI teammate compared to when they work with the low-performing AI teammate. The same results are also found from the two-way ANOVA of human-AI joint performance for *the fair human chess players* and *the poor human chess players*. However, for *the good human*

Table 1
Two-way ANOVA results of human-AI joint performance. Bold indicates statistical significance ($p < 0.05$).

Participants	Source	*F*-Statistic	*p*-value
All Human Chess Players	Teammate Identity	1.207	0.2740
	Teammate Performance	32.36	**8.713×10^{-8}**
	Interaction	0.4346	0.5110
Good Human Chess Players	Teammate Identity	0.01323	0.9092
	Teammate Performance	2.236	0.1460
	Interaction	0.3308	0.5698
Fair Human Chess Players	Teammate Identity	1.275	0.2634
	Teammate Performance	57.47	**2.546×10^{-10}**
	Interaction	0.2341	0.6302
Poor Human Chess Players	Teammate Identity	2.056	0.1627
	Teammate Performance	24.26	**3.403×10^{-5}**
	Interaction	2.056	0.1627

chess players, the effect of teammate performance is not significant ($p = 0.1460$ in Table 1). The means and standard deviations of human-AI joint performance for all participants and participants with different levels of chess expertise are provided in Appendix A.

3.2. Behavioral trust in teammate

Behavioral trust of each participant in their teammate is measured by counting how many times the participant accepts their teammate's move that is different from their initial move as their final move. The results of two-way ANOVA of human behavioral trust in their teammate appear in Table 2. The results of the two-way ANOVA for *all human chess players* suggest that both teammate performance and teammate identity have significant effects on human behavioral trust in their teammate ($p = 8.893 \times 10^{-7}$ and $p = 4.594 \times 10^{-3}$ in Table 2), and the effect of interaction is not significant ($p = 0.6971$ in Table 2). Specifically, for *all human chess players*, they accept their teammate's move that is different from their initial move less often on average when they are deceived about the identity of their AI teammate and believe that they work with another human participant in the study. In addition, human behavioral trust in their teammate becomes higher on average when participants work with the high-performing AI teammate compared to the low-performing AI teammate. The results in Table 2 also indicate that only teammate performance has a significant effect on human behavioral trust in teammate for *the good human chess players* ($p = 7.571 \times 10^{-3}$ in Table 2) and *the fair human chess players* ($p = 9.203 \times 10^{-10}$ in Table 2). In contrast, for *the poor human chess players*, only teammate identity has a significant effect on their behavioral trust in teammate ($p = 8.153 \times 10^{-3}$ in Table 2). The means and standard deviations of human behavioral trust in teammate for all participants and participants with different levels of chess expertise are provided in Appendix B.

3.3. Self-reported trust in teammate

Participants' perceived competency and helpfulness of their teammate reported in the post-study questionnaire represent their self-reported trust in teammate. As shown in Fig. 3, the results of the post-study questionnaire of *all human chess players* show that participants report their low-performing teammate to be more competent and helpful when they are told that they work with another human participant (with deception) rather than an AI teammate (without deception). However, when participants work with the high-performing teammate, teammate identity does not have significant influences on participants' perceived competency and helpfulness of their teammate. The results of *all human chess players* also show that teammate performance has significant effects on participants' perceived competency and helpfulness of their teammate no matter whether they are told that they work with another human participant or an AI teammate. For participants with different levels of chess expertise, it is also found that teammate identity has significant effects on *the poor human chess players'* perceived competency and helpfulness of their teammate when they work with the low-performing teammate ($p = 0.01088$ and $p = 0.02004$ in Appendix C), but such effects are not significant for *the fair human chess players* and *the good human chess players* ($p > 0.05$).

Table 2
Two-way ANOVA results of human behavioral trust in teammate. Bold indicates statistical significance ($p < 0.05$).

Participants	Source	*F*-Statistic	*p*-value
All Human Chess Players	Teammate Identity	8.333	**4.594×10^{-3}**
	Teammate Performance	26.78	**8.893×10^{-7}**
	Interaction	0.1522	0.6971
Good Human Chess Players	Teammate Identity	1.859	0.1836
	Teammate Performance	8.285	**7.571×10^{-3}**
	Interaction	0.2066	0.6530
Fair Human Chess Players	Teammate Identity	1.201	0.2775
	Teammate Performance	52.63	**9.203×10^{-10}**
	Interaction	0.02451	0.8761
Poor Human Chess Players	Teammate Identity	8.111	**8.153×10^{-3}**
	Teammate Performance	0.7573	0.3916
	Interaction	0.05633	0.8141

3.4. Perceived workload

Each participant's perceived workload is measured by the six questions of NASA TLX in the post-study questionnaire. The NASA TLX results of *all human chess players* appear in Fig. 4. These results suggest that participants perceive significantly higher temporal demand when they are told that they work with another human participant (with deception) rather than an AI teammate (without deception). The question about temporal demand in NASA TLX is "how hurried or rushed was the pace of the task?". As shown in Fig. 4, no significant result is found for *all human chess players* on the other five scales of NASA TLX. For participants with different levels of chess expertise, it is also found that *the poor human chess players*' perceived temporal demand is significantly higher when they work with the high-performing "human" teammate compared to when they work with the high-performing AI teammate ($p = 4.462 \times 10^{-4}$ in Appendix D), but such effect is not significant for *the fair human chess players* and *the good human chess players* ($p > 0.05$).

4. Discussion

Based on the three research questions introduced in Section 1, the results of this study indicate that (1) participants accept their AI teammate's decisions less often and do not achieve better human-AI joint performance when they perceive their teammate as another human rather than an AI (i.e., "human" vs. AI). However, participants report that their low-performing teammate is more competent and helpful when they are deceived about the identity of their AI teammate and believe that they work with another human (i.e., low-performing "human" vs. low-performing AI). (2) Teammate performance (high-performing vs. low-performing AI) has significant effects on both human trust in AI and human-AI joint performance. (3) The effects of teammate identity and teammate performance on human trust in AI and human-AI joint performance are different for participants with different levels of expertise (good, fair, and poor human chess players). Several detailed results and the potential reasons behind them are discussed as follows:

First, teammate identity ("human" vs. AI) has a significant effect on participants' behavioral trust in their teammate in this study, where behavioral trust is measured by the number of times when each participant accepts their teammate's move that is different from their initial move in the study. Importantly, all participants work with an AI teammate in the study, and it is found that participants accept their AI teammate's decisions less often when they are deceived and believe that they work with another human participant rather than an AI. The effect of such deception on behavioral trust could be attributed to participants' high expectation of their AI teammate's chess expertise. The findings of prior cognitive studies suggest that people are more likely to cooperate with teammates with expertise and experience (Soll & Larrick, 2009; Van Swol, Paik, & Prahl, 2018). In this study, although participants are not informed about AI accuracy in advance, participants may still expect their AI teammate has better chess expertise than a human because they potentially know that AI has defeated human chess champions, such as IBM Deep Blue defeated human chess world champion Garry Kasparov in 1997 (Hsu, 2002). The effect of such deception on behavioral trust also can be explained by the higher temporal demand participants perceive in the study when they are deceived about the identity of their AI teammate and believe that they are working with another human participant. As stated in Section 3.4, participants who are told that they

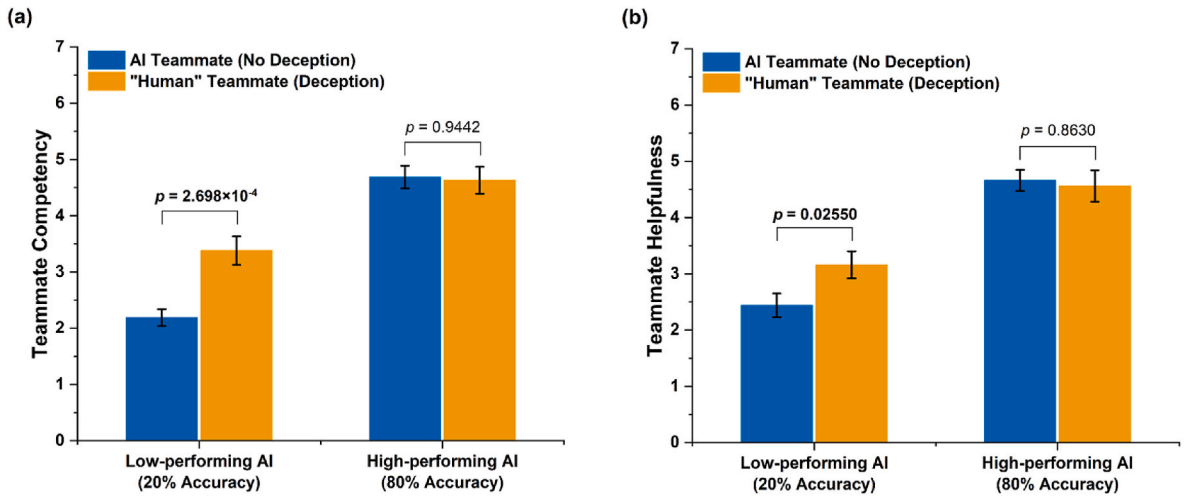

Fig. 3. Participants' perceived competency and helpfulness of their teammate (all human chess players). (a) Teammate competency reported in the post-study questionnaire (1 - "very incompetent", 7 - "very competent"). (b) Teammate helpfulness reported in the post-study questionnarie (1 - "very unhelpful", 7 - "very helpful"). Bold indicates statistical significance ($p < 0.05$).

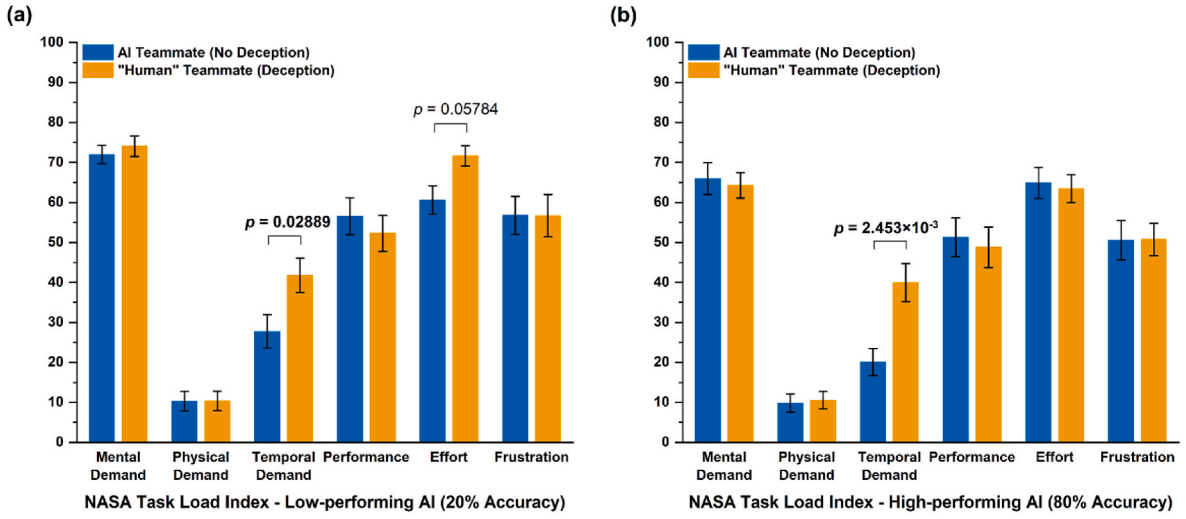

Fig. 4. Participants' perceived workload (all human chess players). (a) Perceived workload when participants work with the low-performing AI teammate (0 - "very low", 100 - "very high"). (b) Perceived workload when participants work with the high-performing AI teammate (0 - "very low", 100 - "very high"). Bold indicates statistical significance ($p < 0.05$).

work with another human participant as their teammate (with deception) report higher temporal demand in the post-study questionnaire than participants who are told that they work with an AI teammate (without deception). Since a relaxed environment is known as a key factor in ensuring effective teamwork (Parker, 2008), participants become less likely to cooperate with their teammate when they feel rushed to complete the task in this study. Here, participants may feel rushed because they perceive that their human teammate keeps waiting for their response in the study.

Second, for participants with different levels of chess expertise, the results in Section 3.2 suggest that only teammate performance (high-performing vs. low-performing AI) has a significant effect on the good human chess players' behavioral trust in their teammate while teammate identity ("human" vs. AI) does not. For the poor human chess players, the opposite is true and only the effect of teammate identity ("human" vs. AI) on behavioral trust is significant. The disparity between the good and poor human chess players is due to the difference of their chess expertise. The good human chess players are better able to judge whether their teammate's move is advantageous in the study than the poor human chess players. This explains why AI teammate performance only exerts a significant impact on the decision making of the good human chess players in the study. In addition, participants' different expectation of their teammate's chess expertise also contributes to the disparity between the good and poor human chess players. Since participants are not given any information about their "human" teammate in the study, such as their teammate's personal information and experience in chess, participants' inference about their "human" teammate's chess expertise is likely based on their own chess expertise (Ross, Greene, & House, 1977; Sherman, Presson, & Chassin, 1984). Thus, the poor human chess players tend to expect a low level of chess expertise for their "human" teammate. Given their potential knowledge about AI's high expertise in chess, the poor human chess players are more likely to cooperate with an AI teammate rather than another human participant in the study. In contrast, the good human chess players tend to reasonably expect a high level of chess expertise for both of their "human" and AI teammates, and teammate identity therefore does not have a significant impact on their decision making in the study.

Third, this study also finds that teammate identity ("human" vs. AI)

has opposite effects on participants' behavioral trust and self-reported trust in their teammate when they work with the low-performing AI teammate. Specifically, participants accept their low-performing teammate's move less often during the study but report that their teammate is more competent and helpful in the post-study questionnaire when they are told that they work with another human participant rather than an AI teammate. The self-reported trust result could be explained by the social pressure participants face when they believe they work with another human participant in the study. Social pressure makes humans act in ways that are consistent with social norms (Hechter & Opp, 2001). In this study, participants may not feel comfortable to report that their human teammate is not competent and helpful because they do not want to offend another human by giving negative feedback. Besides social pressure, participants' expectation of their AI teammate's chess expertise also leads to their lower level of self-reported trust in the low-performing AI teammate. Since participants may expect their AI teammate has a high chess expertise based on their prior knowledge before the study, when the AI turns out as a low-performing teammate during the study, participants are disappointed with the AI and thus give more negative feedback to the AI in the post-study questionnaire.

The results of this study and the potential reasons discussed above provide important implications for real-world applications involving human-AI cooperation, such as human-AI collaborative decision making in the areas of autonomous vehicles, robotic control, manufacturing, and medical services (Contreras-Masse, Ochoa-Zezzatti, García, Pérez-Dominguez, & Elizondo-Cortés, 2020; Ji et al., 2018; Okamura & Yamada, 2020). Specifically, in practice, it is not reasonable to assume that human-AI joint performance can always be improved when humans are deceived into thinking they are working with another human. Our research also warns that deception about the identity of an AI as a human is not similar to the anthropomorphic traits used in prior studies, such as human-like appearance and verbal communication, and such deception may reduce humans' willingness to accept their AI teammate's decisions, especially for humans with a low level of expertise in the task.

This study has several limitations that offer opportunities for future research. First, participants in this study potentially expect that the chess expertise of their AI teammate is better than that of humans. However, people may not have such high expectation of AI's expertise when they work with an AI teammate on other tasks. For example, prior research suggests that people typically have lower expectation of conversational capability for AI than they do for humans (Burgoon et al., 2016; Grimes, Schuetzler, & Giboney, 2021). To generalize the findings of this study, future research needs to study the effects of teammate identity and teammate performance on human-AI cooperation using various tasks in different areas. In addition, participants are not informed about how the AI teammate is created and trained before this study, and the AI teammate also does not explain why each specific move is made during the study. Future research could introduce effective strategies to improve AI transparency and explainability in order to calibrate human trust in AI before and during human-AI cooperation. Moreover, human self-reported trust in their teammate is measured by participants' perceived competency and helpfulness of their teammate at the end of this study. Future research may use other measures related to trust (e.g., participants' self-reported willingness to cooperate with their teammate and participants' self-reported confidence in their teammate) and track the changes of these measures during participants' decision-making process.

5. Conclusions

The results of this study indicate that humans do not achieve better performance when they are deceived and perceive that they work with a human teammate rather than an AI teammate. Moreover, the deception of AI teammate's identity as another human increases human perceived temporal demand and reduces the number of times humans accept their teammate's decision, referred to as human behavioral trust in their teammate, in the study. These results caution against deceiving humans about the identity of their AI teammate in practice because such deception may not be able to enhance human trust in AI and improve human-AI joint performance.

This study also finds that teammate performance (high-performing vs. low-performing AI) and teammate identity ("human" vs. AI) have different effects on humans with different levels of expertise (good, fair, and poor human chess players). Specifically, teammate performance is found to have a significant effect on the good and fair human chess players' behavioral trust in their teammate, but such effect is not significant for the poor human chess players. In contrast, teammate identity only has a significant effect on the poor human chess players' behavioral trust in their teammate in this study. Human individual experience and expertise in a specific task therefore are critical factors that must be considered in future cognitive studies and applications involving human-AI cooperation.

Credit author statement

Guanglu Zhang: Conceptualization, Methodology, Formal analysis, Investigation, Software, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration; **Leah Chong:** Conceptualization, Methodology, Software, Writing - review & editing; **Kenneth Kotovsky:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Funding acquisition; **Jonathan Cagan:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This material is partially supported by the Air Force Office of Scientific Research through cooperative agreement FA9550-18-0088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

Appendix A. Means and standard deviations for human-AI joint performance

Table A.1
Means and standard deviations for human-AI joint performance

Participants	N	Condition	AI Teammate Identity	AI Teammate Performance	Mean	STD
All Human	32	1	No Deception	High-performing AI	33.44	4.21
Chess Players	32	2	No Deception	Low-performing AI	9.28	0.68
	32	3	Deception	High-performing AI	22.19	6.77
	32	4	Deception	Low-performing AI	9.00	0.70
Good Human Chess Players	8	1	No Deception	High-performing AI	51.25	5.43
	8	2	No Deception	Low-performing AI	46.25	5.28
	8	3	Deception	High-performing AI	53.75	5.85
	8	4	Deception	Low-performing AI	42.50	3.42
Fair Human	16	1	No Deception	High-performing AI	35.00	4.92
Chess Players	16	2	No Deception	Low-performing AI	−11.88	5.02
	16	3	Deception	High-performing AI	25.63	5.66
	16	4	Deception	Low-performing AI	−15.63	6.73
Poor Human	8	1	No Deception	High-performing AI	12.50	7.86
Chess Players	8	2	No Deception	Low-performing AI	−51.25	3.28
	8	3	Deception	High-performing AI	−16.25	16.10
	8	4	Deception	Low-performing AI	−51.25	4.49

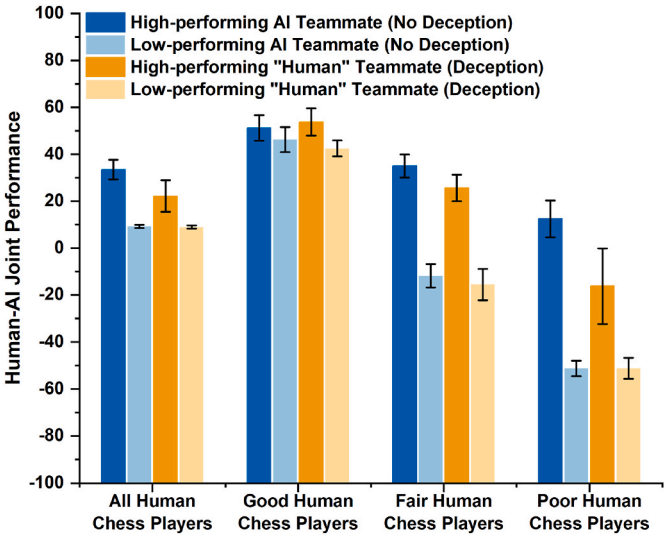

Fig. A.1. Means and standard deviations for human-AI joint performance. Human-AI joint performance of each participant is measured by the participant's cumulative score of the 20 final moves in the study.

Appendix B. Means and standard deviations for human behavioral trust in teammate

Table B.1
Means and standard deviations for human behavioral trust in teammate

Participants	N	Condition	AI Teammate Identity	AI Teammate Performance	Mean	STD
All Human	32	1	No Deception	High-performing AI	7.94	0.58
Chess Players	32	2	No Deception	Low-performing AI	5.06	0.59
	32	3	Deception	High-performing AI	6.44	0.72
	32	4	Deception	Low-performing AI	3.09	0.44
Good Human Chess Players	8	1	No Deception	High-performing AI	4.88	0.94
	8	2	No Deception	Low-performing AI	2.88	0.62
	8	3	Deception	High-performing AI	4.13	0.96
	8	4	Deception	Low-performing AI	1.38	0.43
Fair Human	16	1	No Deception	High-performing AI	8.63	0.60
Chess Players	16	2	No Deception	Low-performing AI	4.38	0.40
	16	3	Deception	High-performing AI	8.06	0.72
	16	4	Deception	Low-performing AI	3.63	0.55
Poor Human	8	1	No Deception	High-performing AI	9.63	1.21
Chess Players	8	2	No Deception	Low-performing AI	8.63	1.50
	8	3	Deception	High-performing AI	5.50	1.94
	8	4	Deception	Low-performing AI	3.75	1.11

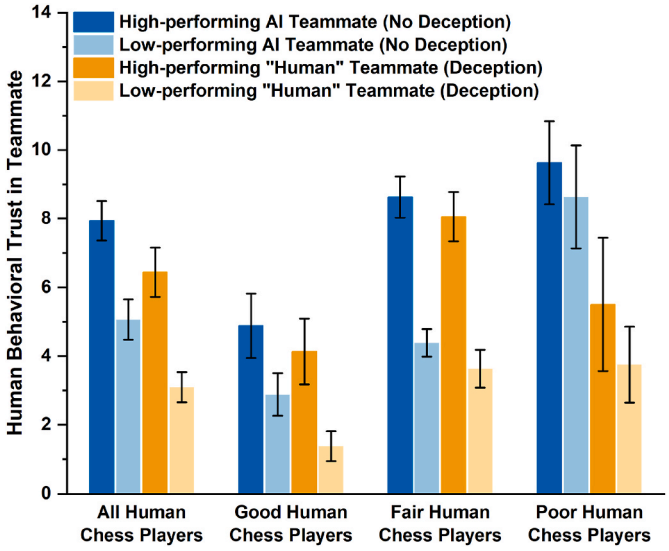

Fig. B.1. Means and standard deviations for human behavioral trust in teammate. Behavioral trust of each participant in their teammate is measured by counting how many times the participant accepts their teammate's move that is different from their initial move as their final move.

Appendix C. The poor human chess players' self-reported trust in teammate

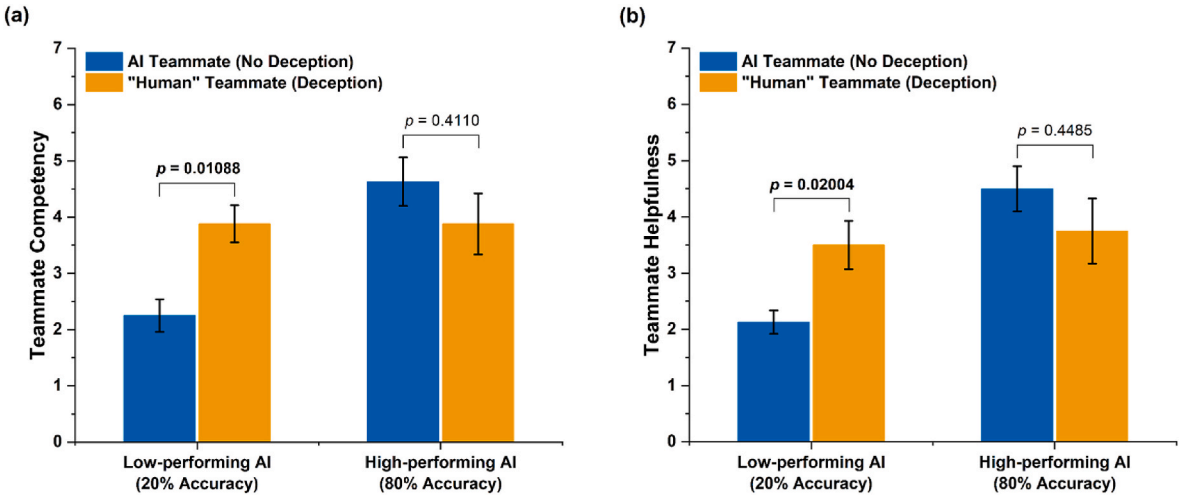

Fig. C.1. The poor human chess players' perceived competency and helpfulness of their teammate. (a) Teammate competency reported in the post-study questionnaire (1 - "very incompetent", 7 - "very competent"). (b) Teammate helpfulness reported in the post-study questionnarie (1 - "very unhelpful", 7 - "very helpful"). Bold indicates statistical significance ($p < 0.05$).

Appendix D. The poor human chess players' perceived workload

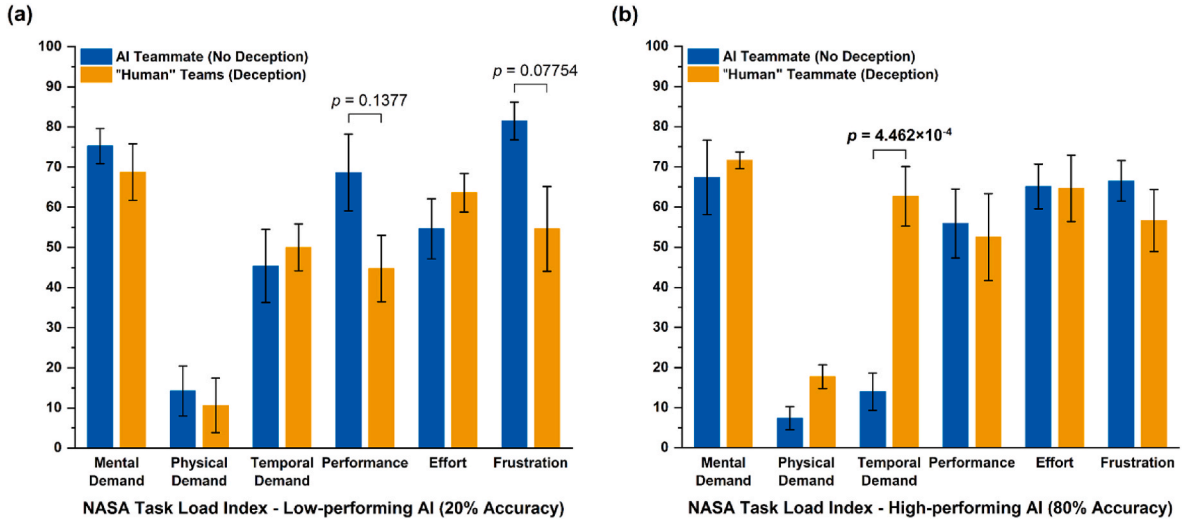

Fig. D.1. The poor human chess players' perceived workload. (a) Perceived workload when participants work with the low-performing AI teammate (0 - "very low", 100 - "very high"). (b) Perceived workload when participants work with the high-performing AI teammate (0 - "very low", 100 - "very high"). Bold indicates statistical significance ($p < 0.05$).

References

Bao, L., Krause, N. M., Calice, M. N., Scheufele, D. A., Wirz, C. D., Brossard, D., et al. (2022). Whose AI? How different publics think about AI and its social impacts. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2022.107182, 107182.

Beck, J., Stern, M., & Haugsjaa, E. (1996). Applications of AI in education. *XRDS: Crossroads, The ACM Magazine for Students, 3*(1), 11–15. https://doi.org/10.1145/332148.332153

Boone, R. T., & Buck, R. (2003). Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior, 27*(3), 163–182. https://doi.org/10.1023/A:1025341931128

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., et al. (2016). Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies, 91*, 24–36. https://doi.org/10.1016/j.ijhcs.2016.02.002

Carson, T. L. (2010). *Lying and deception: Theory and practice*. Oxford University Press.

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior, 127*, Article 107018. https://doi.org/10.1016/j.chb.2021.107018

Contreras-Masse, R., Ochoa-Zezzatti, A., García, V., Pérez-Dominguez, L., & Elizondo-Cortés, M. (2020). Implementing a novel use of multicriteria decision analysis to select IIoT platforms for smart manufacturing. *Symmetry, 12*(3), 368. https://doi.org/10.3390/sym12030368

Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior, 29*(3), 577–579. https://doi.org/10.1016/j.chb.2012.11.023

Dujmovic, J. (2017). Opinion: What's holding back artificial intelligence? Americans don't trust it. MarketWatch. Retrieved from https://www.marketwatch.com/story/whats-holding-back-artificial-intelligence-americans-dont-trust-it-2017-03-30. (Accessed 7 July 2022) Accessed.

Fox, J., Ahn, S. J., Janssen, J. H., Yeykelis, L., Segovia, K. Y., & Bailenson, J. N. (2015). Avatars versus agents: A meta-analysis quantifying the effect of agency on social influence. *Human-Computer Interaction, 30*(5), 401–432. https://doi.org/10.1080/07370024.2014.921494

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems, 144*, Article 113515. https://doi.org/10.1016/j.dss.2021.113515

Gyory, J. T., Soria Zurita, N. F., Martin, J., Balon, C., McComb, C., Kotovsky, K., et al. (2022). Human versus artificial intelligence: A data-driven approach to real-time process management during complex engineering design. *Journal of Mechanical Design, 144*(2), Article 021405. https://doi.org/10.1115/1.4052488

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hechter, M., & Opp, K.-D. (2001). *Social norms*. Russell Sage Foundation.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hsu, F.-H. (2002). *Behind Deep Blue: Building the computer that defeated the world chess champion*. Princeton University Press.

Jha, S. K., Bilalovic, J., Jha, A., Patel, N., & Zhang, H. (2017). Renewable energy: Present research and future scope of artificial intelligence. *Renewable and Sustainable Energy Reviews, 77*, 297–317. https://doi.org/10.1016/j.rser.2017.04.018

Ji, P., Zeng, H., Song, A., Yi, P., Xiong, P., & Li, H. (2018). Virtual exoskeleton-driven uncalibrated visual servoing control for mobile robotic manipulators based on human-robot-robot cooperation. *Transactions of the Institute of Measurement and Control, 40*(14), 4046–4062. https://doi.org/10.1177/0142331217741538

Kulms, P., & Kopp, S. (2019). More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation. In *Proceedings of Mensch Und Computer* (pp. 31–42). https://doi.org/10.1145/3340764.3340793, 2019 Hamburg, Germany.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

McNeese, N. J., Schelble, B. G., Canonico, L. B., & Demir, M. (2021). Who/what is my teammate? Team composition considerations in human-AI teaming. *IEEE Transactions on Human-Machine Systems, 51*(4), 288–299. https://doi.org/10.1109/THMS.2021.3086018

Metz, C. (2021). Who is making sure the AI machines aren't racist. The New York Times. . Retrieved from https://www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html. (Accessed 7 July 2022).

O'Leary, D. E. (2019). Google's Duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management, 26*(1), 46–53. https://doi.org/10.1002/isaf.1443

Okamura, K., & Yamada, S. (2020). Empirical evaluations of framework for adaptive trust calibration in human-AI cooperation. *IEEE Access, 8*, 220335–220351. https://doi.org/10.1109/ACCESS.2020.3042556

Parker, G. M. (2008). *Team players and teamwork: New strategies for developing successful collaboration*. John Wiley & Sons.

Pazzanese, C. (2020). Ethical concerns mount as AI takes bigger decision-making role in more industries. The Harvard Gazette. . Retrieved from https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/. (Accessed 7 July 2022).

Pelau, C., Dabija, D.-C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior, 122*, Article 106855. https://doi.org/10.1016/j.chb.2021.106855

Plastino, E. (2021). Data modernization: Breaking the AI vicious cycle for superior decision-making. In *The cognizant's center for the future of work*New Jersey, USA: Teaneck. Whitepaper retrieved from https://www.cognizant.com/futureofwork/whitepaper/data-modernization-breaking-the-ai-vicious-cycle-for-superior-decision-making. (Accessed 7 July 2022) Accessed.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*(3), 279–301. https://doi.org/10.1016/0022-1031(77)90049-X

Seeber, I., Bittner, E., Briggs, R. O., De Vreede, T., De Vreede, G.-J., Elkins, A., et al. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management, 57*(2), Article 103174. https://doi.org/10.1016/j.im.2019.103174

Sergio. (2022). *Stockfish wins TCEC Season 22, sets records*. Retrieved from. CHESSDOM Accessed July 7, 2022 https://www.chessdom.com/stockfish-wins-tcec-season-22-sets-records/.

Sherman, S. J., Presson, C. C., & Chassin, L. (1984). Mechanisms underlying the false consensus effect: The special role of threats to the self. *Personality and Social Psychology Bulletin, 10*(1), 127–138. https://doi.org/10.1177/0146167284101015

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. Retrieved from *Cutter Business Technology Journal, 31*(2), 47–53 Accessed July 7, 2022. https://www.cutter.com/article/building-trust-artificial-intelligence-machine-learning-and-robotics-498981.

Simon, H. A., Bibel, W., Bundy, A., Berliner, H., Feigenbaum, E. A., Buchanan, B. G., et al. (2000). AI's greatest trends and controversies. *IEEE Intelligent Systems and Their Applications, 15*(1), 8–17. https://doi.org/10.1109/5254.820322

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 780–805. https://doi.org/10.1037/a0015145

Sujan, M., Baber, C., Salmon, P., Pool, R., & Chozos, N. (2021). *Human factors and ergonomics in healthcare AI. The chartered institute of ergonomics and human factors*. Wootton Park, UK. Whitepaper retrieved from. . Accessed July 7, 2022 https://ergonomics.org.uk/resource/human-factors-in-healthcare-ai.html.

Turing, A. M. (2009). Computing machinery and intelligence. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the turing test* (pp. 23–65). Springer. https://doi.org/10.1007/978-1-4020-6710-5_3.

Van Dongen, K., & Van Maanen, P.-P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies, 71*(4), 410–424. https://doi.org/10.1016/j.ijhcs.2012.10.018

Van Swol, L. M., Paik, J. E., & Prahl, A. (2018). Advice recipients: The psychology of advice utilization. In E. L. MacGeorge, & L. M. Van Swol (Eds.), *The Oxford handbook of advice* (pp. 21–41). Oxford University Press.

Verberne, F. M., Ham, J., & Midden, C. J. (2015). Trusting a virtual driver that looks, acts, and thinks like you. *Human Factors, 57*(5), 895–909. https://doi.org/10.1177/0018720815580749

de Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., et al. (2012). The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56* pp. 263–267). https://doi.org/10.1177/1071181312561062. Los Angeles, CA.

de Visser, E. J., Monfort, S. S., Goodyear, K., Lu, L., O'Hara, M., Lee, M. R., et al. (2017). A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. *Human Factors, 59*(1), 116–133. https://doi.org/10.1177/0018720816687205

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., et al. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*(3), 331–349. https://doi.org/10.1037/xap0000092

Von der Pütten, A. M., Krämer, N. C., Gratch, J., & Kang, S.-H. (2010). It doesn't matter what you are!" Explaining social effects of agents and avatars. *Computers in Human Behavior, 26*(6), 1641–1650. https://doi.org/10.1016/j.chb.2010.06.012

Warwick, K., & Shah, H. (2016). Can machines think? A report on turing test experiments at the royal society. *Journal of Experimental & Theoretical Artificial Intelligence, 28*(6), 989–1007. https://doi.org/10.1080/0952813X.2015.1055826

Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. Retrieved from *Harvard Business Review, 96*(4), 114–123 Accessed July 7, 2022 https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces.

Zhang, G., Raina, A., Cagan, J., & McComb, C. (2021). A cautionary tale about the impact of AI on human design teams. *Design Studies, 72*, Article 100990. https://doi.org/10.1016/j.destud.2021.100990