Computers in Human Behavior 127 (2022) 107018

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice

Leah Chong [a], Guanglu Zhang [a], Kosa Goucher-Lambert [b], Kenneth Kotovsky [c], Jonathan Cagan [a,*]

[a] Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[b] Department of Mechanical Engineering, University of California Berkeley, Berkeley, CA, 94720, USA
[c] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Human-AI interaction
Decision-making
Confidence
Trust

ABSTRACT

Artificial intelligence (AI) has shown its promise in assisting human decision-making. However, humans' inappropriate decision to accept or reject suggestions from AI can lead to severe consequences in high-stakes AI-assisted decision-making scenarios. This problem persists due to insufficient understanding of human trust in AI. Therefore, this research studies how two types of human confidence that affect trust, their confidence in AI and confidence in themselves, evolve and affect humans' decisions. A cognitive study and a quantitative model together examine how changing positive and negative experiences affect these confidences and ultimate decisions. Results show that human self-confidence, *not* their confidence in AI, directs the decision to accept or reject AI suggestions. Furthermore, this work finds that humans often misattribute blame to themselves and enter a vicious cycle of relying on a poorly performing AI. Findings reveal the need and provide insights to effectively calibrate human self-confidence for successful AI-assisted decision-making.

1. Introduction

The promise of artificial intelligence (AI) systems to pull insights from large data has led them to be used in collaboration with humans in various decision-making domains involving healthcare, business, military, and design (Buch et al., 2018; Kamar et al., 2012; Nagar & Malone, 2011; Parasuraman et al., 2009; Patel et al., 2019; Zhang et al., 2021). Human-AI teams are increasingly deployed to improve joint performance and accomplish tasks that neither an AI nor human can solve alone (Wilson & Daugherty, 2018). Humans often remain responsible for the final decisions due to ethical concerns; therefore, these teams can only reach their collaborative potential when human decision-makers appropriately accept or reject AI input (Zhang et al., 2020).

However, humans are prone to error. Their failure to discern when to accept or reject AI input hinders team performance. This is especially problematic in high-stakes situations where decisions affect human lives such as self-driving cars and medical diagnosis (Lee & See, 2004; Parasuraman & Riley, 1997; Zhang et al., 2020). For example, in the 2015 Google car crash, the Google car was hit from behind when the driver manually applied the brakes while the car was slowing for a pedestrian (Richtel & Dougherty, 2015). Although this crash only resulted in a mild whiplash, human's inappropriate judgement can lead to more severe accidents. Not only this, the performance of Watson for Oncology, IBM's cancer treatment recommendation system, varies greatly depending on the population and the type of cancer (Strickland, 2019). If doctors fail to reject Watson's faulty recommendations, patients can receive inappropriate treatment for their cancer.

Recent literature points to inappropriate trust as the reason for under- or over-relying on AI (Bansal et al., 2019a, 2019b, Dzindolet et al., 2003; Hoffman et al., 2013; Lee & See, 2004; Parasuraman & Riley, 1997; Siau & Wang, 2018; Zhang et al., 2020), meaning that humans accept or reject AI suggestions when they should not because their trust for the AI does not match the AI's trustworthiness. Despite the AI capabilities that sometimes outperform human judgment, many people still hesitate to trust AI. According to recent surveys, 42% of participants lack general trust in AI, and 49% of participants could not name any AI product they trust (Dujmovic, 2017). Furthermore, humans are highly unforgiving of AI error, subsequently distrusting AI input regardless of its quality (Alvarado-Valencia & Barrero, 2014; Dietvorst et al., 2015). However, in some cases, people over-trust AI. One major

* Corresponding author. Department of Mechanical Engineering, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA, 15213, USA.
E-mail address: cagan@cmu.edu (J. Cagan).

https://doi.org/10.1016/j.chb.2021.107018
Received 27 February 2021; Received in revised form 29 June 2021; Accepted 7 September 2021
Available online 10 September 2021

source of this over-trust is high cognitive load in solving complex tasks where people increasingly desire to rely on AI and overestimate AI capabilities (Alvarado-Valencia & Barrero, 2014; Goddard et al., 2012; Parasuraman et al., 1993; Zhang et al., 2021).

Despite the increasing attention, human trust in AI has not been understood enough to avoid faulty decisions to accept or reject AI input. This understanding has been difficult because of the diverse and abstract definitions of trust (Dietz & Den Hartog, 2006; Jonker & Treur, 1999; Lewicki et al., 2006; Lewis & Weigert, 1985; Mcknight & Chervany, 1996). Thus, instead of directly studying trust, this work focuses on two types of human confidence that are often discussed alongside trust: confidence in AI and self-confidence. Confidence in AI is formed from trustor's perception of trustee's (in this case an AI) ability to perform a given task, while self-confidence contributes to the trustor's willingness to rely on trustee. Confidence in AI and self-confidence respectively provide insight into two antecedents of trust proposed by Mayer et al.: perception of trustee's attributes such as ability, and a propensity to trust (related to a personal disposition to trust) (Mayer et al., 1995; Rousseau et al., 1998). Knowledge about how the two types of confidence evolve as a result of different experiences (e.g., humans accept/reject AI suggestions and then receive positive/negative feedback) can prompt ways to prevent inappropriate acceptance or rejection of AI input by influencing human confidence.

Prior works related to confidence in AI and self-confidence are typically limited to considering constant levels of confidence, not accounting for the highly complex and dynamic nature of human trust in AI (Crisp & Jarvenpaa, 2013; Danks & London, 2017; Glikson & Woolley, 2020; Mayer et al., 1995; Schoorman et al., 2007; Yin et al., 2019); confidence in AI and self-confidence alter as the AI's abilities and performance vary based on its capabilities, available data to train the AI, and problem-solving situations. An accurate understanding of confidence dynamics is critical to effectively reduce erroneous reliance on AI and improve human-AI team performance.

In order to close this gap in knowledge, this work investigates the evolution of human confidence in AI and in themselves, and their impact on human-AI decision-making. Specifically, this study explores one of the most prevalent collaboration contexts known as AI-assisted decision-making, where humans are responsible for the final decision; after receiving a suggestion from their AI teammate, humans either accept or override the suggestion. In this work, AI is functioning as a "second opinion" system where it provides its solution to humans as a second opinion to consider. This work presents an empirical study and establishes a quantitative model to explore the following research questions: 1) how do changes in AI performance and resulting positive and negative feedback affect human confidence in AI and human self-confidence? 2) how are these two types of confidence associated with the probability of accepting AI suggestions? and 3) what decision-making pattern distinguishes those who successfully accept and reject AI suggestions?

2. Methods

A human subject study and a quantitative model are used to understand human cognition and AI acceptance behavior during an AI-assisted decision-making scenario. The cognitive study is designed to reveal the real-time variation in human confidence in AI and human self-confidence during AI-assisted decision-making as a result of AI performance change. The quantitative model of human confidence is developed to capture the impact of different experiences during AI-assisted decision-making on confidence dynamics.

2.1. Human subject study

2.1.1. Experimental task

Participants are given a chess puzzle task in which they, in collaboration with an AI, must make the best chess move given a chess board state (see Fig. 1). This task enables a wide range of possible decision choices and the unpredictability of the final outcome resulting from the current decision, representing many decision-making scenarios in the real-world. Furthermore, the availability of an open-source chess engine, Stockfish (https://github.com/official-stockfish/Stockfish), that can perform above expert human level, makes the chess puzzle task an excellent fit for this study. Though the participants are aware that they are working with an AI, the AI is given a gender-neutral name, Taylor, to reduce any gender bias.

Because of its capabilities, Stockfish is used to design the AI teammate for the experiment. Given a chess board state, Stockfish uses a minimax search tree to look for a list of possible moves. Each move suggested by Stockfish has an evaluation score attached to it which represents how advantageous or disadvantageous the move is, given the current board state.

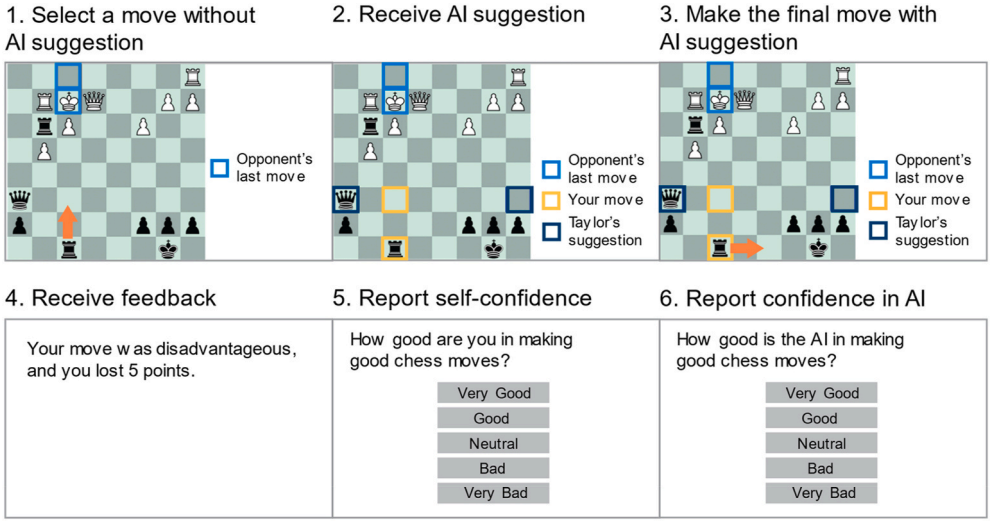

Fig. 1. Task procedure. The figure demonstrates the step-by-step procedure of the experiment. Participants initially select a move without AI suggestion (orange arrow in the first screen image), receive an AI (i.e., Taylor) suggestion (dark blue boxes), and make their final move (orange arrow in the third image). The final move can be the AI suggestion or any other move. After receiving feedback on their final move, participants report their self-confidence and their confidence in the AI.

2.1.2. Participants

100 participants are recruited for (and completed) the experiment in accordance with a protocol approved by the University's Institutional Review Board. All participants are fluent English speakers and know how to play chess prior to their participation. Informed consent is obtained from all participants before inclusion.

2.1.3. Experimental conditions

There are two experimental condition groups: the positive and negative AI performance change groups. Participants are randomly assigned to these groups, 50 in each group. The two conditions both include two AI performance levels: 80% and 20% accuracy, with the participants experiencing the two levels in opposite orders. In Condition 1, AI performance changes from 80% to 20% accuracy, while in Condition 2, it changes from 20% to 80% accuracy. When the AI is 80% accurate, for 80% of the time it chooses the best move out of the list of moves Stockfish provides. The best move always has a positive evaluation score, meaning that it is an advantageous move. For the other 20% of the times, the AI chooses the 7th best move which is ensured to always have a negative evaluation score (i.e., a disadvantageous move).

The chess puzzles in the study are chosen by considering their difficulty and possible next moves. The publicly available collection of Mate-in-4 board states (http://wtharvey.com/m8n4.txt) are first explored using the Stockfish engine. Mate-in-4 board states are used to ensure that the game has not gotten too close to the end where participants can easily see and pick the best move. With the Mate-in-4 board states, there is a wide range of possible moves, and the final outcome remains unpredictable. For a given board state, the possible moves and their corresponding "goodness" evaluation scores are found. Only the board states with at least two advantageous and two disadvantageous moves are included in the experiment to keep a fairly constant puzzle difficulty. It is important to note that although the Mate-in-4 board states are chosen to ensure problem consistency, the task in this experiment is to make a single move (not 4) with an AI agent. The description of the experimental procedure can be found in Section 2.1.4.

Following 3 practice chess puzzles, participants in both conditions solve 30 puzzles to capture repetitive collaboration scenarios and allow for an examination of how individual experience impacts human confidence in the AI and in themselves. After the first 20 puzzles, the AI performance level changes, instantiating the dynamic performance of the AI.

2.1.4. Procedure

The experiment is conducted via Amazon Web Services. Before the experiment, all participants are asked to sign an informed consent using Google Forms. Then, the participants are provided with the step-by-step instruction of the experiment via email. During their 90-min time slot, the participants follow the instructions to complete the experiment.

The participants perform the following procedure (see Fig. 1) for each puzzle. First, the participants are asked to select their best move without the AI suggestion. This is to collect data about their individual chess skill without the AI's help. Once the participants make the unassisted selection, they receive the AI suggestion and are asked to make the final decision to accept or override it. Here, the study is designed so that when overriding the AI suggestion, the participants are not limited to their unassisted selection but can make any move different from the AI suggestion. Next, the participants gain and lose 5 points according to the feedback (i.e., advantageous or disadvantageous) they receive on their final move. The feedback is given based on the evaluation score computed by the Stockfish engine. In the beginning of the experiment, the participants are informed about the scoring system and that those who receive a score of 40 or above at the end of the experiment will receive an additional monetary prize. Finally, the participants are asked to report their confidence in their own ability and the AI's ability in solving chess puzzles in a 5-point Likert scale. The confidence questions ask: How good are you (or the AI) in making good chess moves? The 5-point Likert scale includes answers: very good, good, neutral, bad, and very bad, which are quantified as 1, 0.75, 0.5, 0.25, and 0 for model fitting.

2.2. Confidence model

2.2.1. Concept and structure

Fig. 2 illustrates the concept structure of the model. The model predicts the dynamics of human confidence based on experience, accumulated confidence, and bias, inspired by Hu et al.'s dynamic trust model (Hu et al., 2019). Our confidence model is applied to both human confidence in AI and self-confidence. Depending on the object of confidence (AI or self), confidence, accumulated confidence, and bias terms change accordingly, while the experience term stays the same as it is not conditional on the object.

Hu et al.'s model is built for a different type of human-AI collaboration where the AI provides specific information about the problem-solving environment from which humans decide to accept or reject that information; the humans cannot look for the information themselves and are completely dependent on what the AI reports. In contrast, our model represents a more complex yet common scenario where humans accept or override the AI's suggestion as humans and AI are both working on the same problem-solving task. This means that rather than simply deciding to accept or reject the AI report, humans receive the same amount of information about the problem as the AI does. The AI is therefore functioning as a "second opinion" system in which it provides its solution as a "second opinion" for humans to consider. Then, humans decide to accept the AI suggestion or alternatively, override the AI suggestion to replace with their own solution. Therefore, the proposed confidence model extends Hu et al.'s model, including an entirely different set of possible experiences in the experience term ($E(n)$). Additionally, the current model extends to predict not only the confidence in AI but also the confidence in themselves.

2.2.2. Model description

The general model equation is as follows (Hu et al., 2019):

$$C(n+1) = C(n) + \alpha_e[E(n) - C(n)] + \alpha_a[A(n) - C(n)] + \alpha_b[B(n) - C(n)], \tag{1}$$

where $C(n)$, $E(n)$, $A(n)$, $B(n)$, α_e, α_a, $\alpha_b \in [0,1]$.

The three factors of confidence at trial n: experience ($E(n)$), accumulated confidence ($A(n)$), and bias ($B(n)$), are compared to the confidence at trial n ($C(n)$) to deduce how each factor would affect current confidence. Then the weighted sum of the three difference values (i.e., $E(n) - C(n)$, $A(n) - C(n)$, and $B(n) - C(n)$) is added to the current confidence to calculate the confidence in trial $n+1$. The weights α_e, α_a, and α_a are the rate factors.

This work mainly explores the model's experience term. Experience refers to the experience humans have with the AI at trial n. In AI-assisted decision-making contexts, humans can have one of the following four experiences in each trial:

1) accept the AI suggestion, then receive positive feedback ($e_1(n)$);
2) reject the AI suggestion, then receive positive feedback ($e_2(n)$);
3) accept the AI suggestion, then receive negative feedback ($e_3(n)$);
4) reject the AI suggestion, then receive negative feedback ($e_4(n)$).

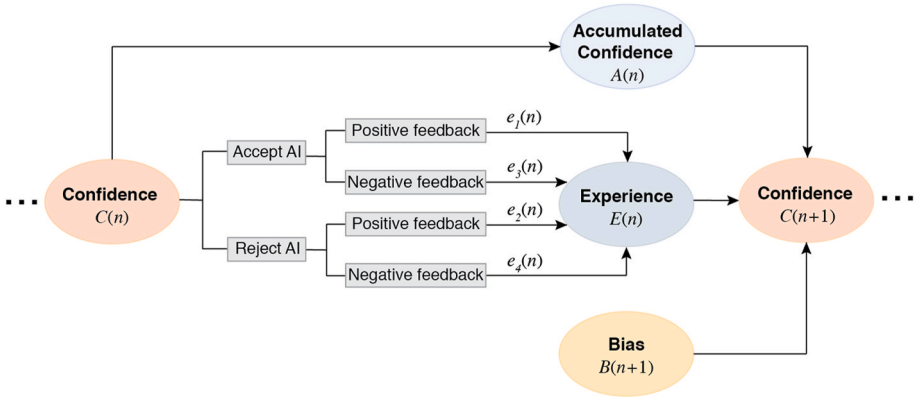

Fig. 2. Concept structure of the confidence model. The model includes three major factors of human confidence: experience, accumulated confidence, and bias. The experience term ($E(n)$) considers four different possible experiences humans can have with an AI at trial n. The accumulated confidence term ($A(n)$) represents the confidence levels leading up to the current trial n. The bias term ($B(n)$) is the inherent bias towards any AI system or towards themselves at trial n. The confidence value of the next trial is computed using these three factors.

Therefore, the experience term, $E(n)$, in Eq. (1) is quantitatively represented as the sum of these four sub-experience terms with their respective weights (ω_1, ω_2, ω_3, and ω_4), which are referred to as the experience impact factors:

$$E(n) = \omega_1 e_1(n) + \omega_2 e_2(n) + \omega_3 e_3(n) + \omega_4 e_4(n), \quad (2)$$

where $e_1(n)$, $e_2(n)$, $e_3(n)$, $e_4(n) = 0$ *or* 1,

ω_1, ω_2, ω_3, $\omega_4 \in [0, 1]$,

and $\sum_{n=1}^{4} e_n = 1$.

The values of these sub-experience terms are binary (i.e., 0 or 1) for each trial n because each type of experience either happens (i.e., 1) or does not happen (i.e., 0). Additionally, the sum of these terms at a trial equals to 1, meaning only one of these experiences occurs per trial (e.g., accepting the AI suggestion then receiving positive feedback is represented by $e_1(n)$=1, $e_2(n)$=0, $e_3(n)$=0, and $e_4(n)$=0).

Accumulated confidence is the confidence levels from the previous trials that are accumulated into a value with the time discounting factor γ (Hu et al., 2019):

$$A(n) = \gamma C(n-1) + (1-\gamma)A(n-1), \quad (3)$$

where $\gamma \in [0, 1]$,

and $A(0) = C(0)$.

This time discounting factor in the range [0, 1] accounts for the possible recency bias in which the confidence values from newer trials are considered to be more impactful than those from the older trials. The accumulated confidence values at trial 0 and 1 (i.e., $A(0)$ and $A(1)$) are equal to the initial confidence because there have not been any updates in the confidence yet.

Bias represents the general bias for AI systems, not for a specific AI (Hu et al., 2019). For this experiment, this term is assumed to be constant for simplicity and set to the initial confidence value before the experiment starts:

$$B(n) = B(n-1), \quad (4)$$

where $B(0) = C(0)$.

2.2.3. Parameter fitting

There are 8 parameters in the confidence model: α_e, α_a, α_b, ω_1, ω_2, ω_3, ω_4, and γ. The optimal parameter values (see Table 1) that minimize the least squares error of the fit are estimated with the experimental data from all trials in the cognitive study. Trust region reflective algorithm (Moré & Sorensen, 1983) is employed for parameter estimation. Because nonlinear regression is sensitive to the initial guess, the parameter estimation process is repeated with varying initial values, confirming the estimated parameter values to be consistent. Using the estimated parameter values and the initial confidence values, the model iteratively computes the human confidence in AI and human self-confidence values for each trial in the experiment (see Fig. 3). The mean squared error (MSE) and adjusted R-squared value of the confidence in AI model are 0.0020 and 0.700 respectively, while those of the self-confidence model are 0.0012 and 0.478. The distinct R-squared values suggest that there might be differences between confidence in AI and self-confidence that are not captured fully by the model. A possible difference is that people may have more variability in perceiving and reporting their self-confidence, as opposed to their confidence in the AI.

A robustness test is conducted to ensure that the model parameter fitting results in Table 1 are robust. 80% of the experimental data (80 participants) are randomly selected 100 times. Then, the parameters are estimated using each of the chosen 80% data. The initial parameter values are set as shown in Table 1. The results of the robustness test show that the estimated parameters from the 100 trials are consistent with minor differences (below 0.05 mean absolute deviation from the fitting results in Table 1), confirming that the model parameter fitting results are robust. Not only this, notably, confidence in AI and self-confidence are in the range of [0,1] in the experiment, which leads to small MSE values.

3. Results

3.1. Impact of dynamic AI performance on human confidence in AI and in themselves

Fig. 4 illustrates plots of human confidence in the AI and human self-confidence from two experimental conditions, with varied and dynamic AI performers advising chess moves for different board positions (i.e., puzzles): Condition 1 where a good performing AI changes to a poor

Table 1
Model parameter fitting results. The quantitative model of confidence in AI and self-confidence are fitted to the experimental data.α_e, α_a, α_b are the rate factors of experience, accumulated confidence, and bias, respectively. ω_1, ω_2, ω_3, $\omega_4$are the experience impact factors of the four type of experiences in AI-assisted decision-making. γ is the time discounting factor.

	α_e	α_a	α_b	ω_1	ω_2	ω_3	ω_4	γ
Confidence in AI	0.2672	0.3405	0.05240	0.8439	0.2115	~0	0.5217	0.3897
Self-confidence	0.2844	0.4706	~0	0.5736	0.8284	0.2384	0.2863	0.1147

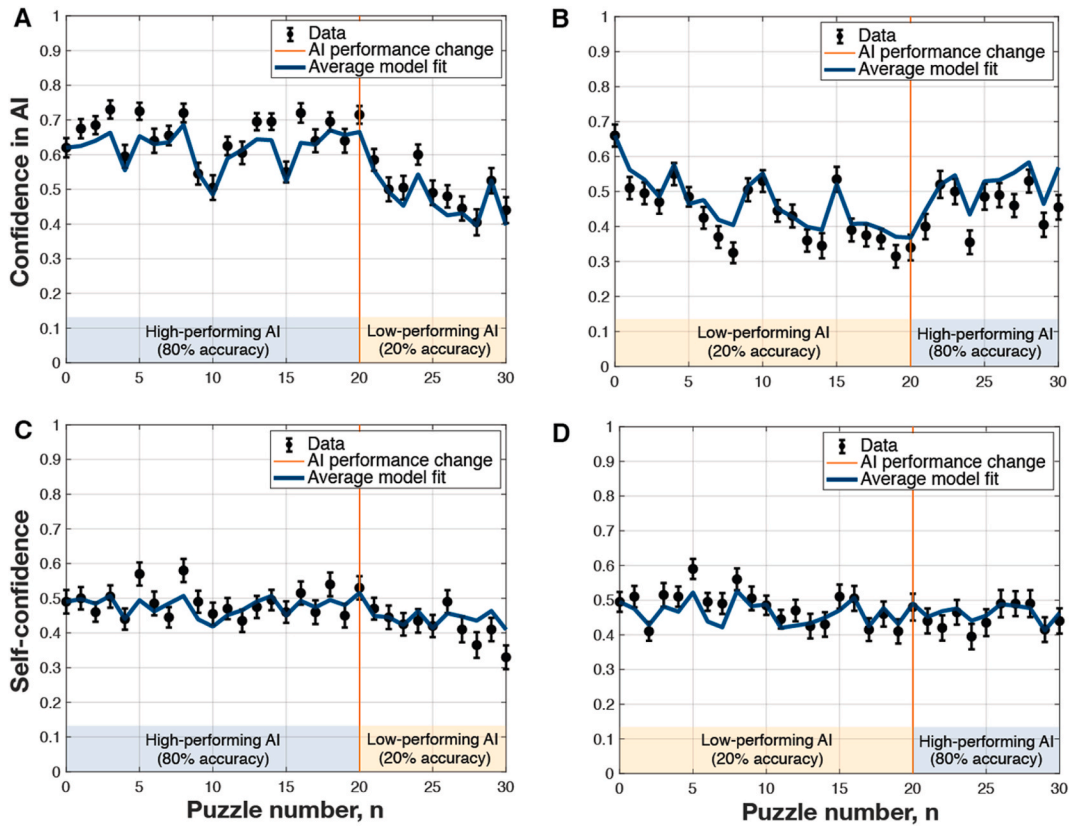

Fig. 3. Confidence in AI and self-confidence model fitting results in each experimental condition. (A) and (B) are plots of human confidence in the AI in Conditions 1 and 2 respectively. (C) and (D) are plots of human self-confidence in Conditions 1 and 2 respectively. Each pair (i.e., (A) and (B), and (C) and (D)) uses the same parameter values to fit the model to the data. The black data points are the average respective confidence values of the 50 participants in each condition. The error bars represent the standard error of the data. The dark blue lines show the average model fitting results with the estimated parameter values among the 50 participants in each trial.

performing one two-thirds of the way through the study (i.e., after puzzle 20), and Condition 2 where a poor performing AI changes to a good performing one. The participants are uninformed of this performance change. The linear fits of the average confidence data before and after the AI performance change are used for analysis.

Fig. 4A and B show that when initially interacting with the AI from puzzles 1 to 20, good AI performance does not significantly affect the human confidence in the AI (F-test, P=0.9), while poor performance decreases it (F-test, P<0.001). After the switch in the AI performance after puzzle 20, the trend in the confidence in the AI (i.e., dark blue line) alters significantly in both conditions in the same direction as the AI performance change. Specifically, the slope changes in the negative direction in Condition 1 (linear regression with interaction, P<0.05) and in the positive direction in Condition 2 (linear regression with interaction, P<0.05).

Fig. 4C and D demonstrate that when initially interacting with the AI from puzzles 1 to 20, both conditions do not show a significant impact on self-confidence, meaning neither good nor bad performance by the AI affects self-confidence (F-test, P=1 and 0.06). When the behavior of the AI changes, the impact of this change on self-confidence is different depending on the direction of change. When the AI changes to perform worse (i.e., Condition 1), self-confidence shows a decreasing trend (linear regression with interaction, P<0.05). Yet, when the AI changes for the better (i.e., Condition 2), self-confidence is not affected significantly (linear regression with interaction, P=0.4). Therefore, the participants' self-confidence changes significantly only when an initially good-performing AI starts to perform badly.

3.2. *Impact of different types of experiences on human confidence in AI and in themselves*

The model parameters are estimated from the experimental data to evaluate the impact of the four types of experiences during AI-assisted decision-making on human confidence in the AI and in themselves. These experiences include receiving either positive or negative feedback on the performance of the AI (i.e., accept the AI suggestion) and those on their own performance (i.e., reject the AI suggestion). The estimated values of parameters, ω_1, ω_2, ω_3, and ω_4, represent the impact of these experiences. These values range from 0 to 1, 0.5 meaning no or neutral impact; values below 0.5 mean negative impact, and those above 0.5 mean positive impact.

The first row of parameter values in Table 2 shows how the four types of experiences affect the participants' confidence in the AI. The parameter values of ω_1 and ω_3 convey that positive and negative feedback on the performance of the AI strongly influence the participants' confidence in the AI in the expected direction (ω_1=0.844 and ω_3=0.000). Particularly, negative feedback on the AI performance greatly decreases participants' confidence in the AI. However, when receiving feedback on their own move, positive feedback decreases the participants' confidence in the AI (ω_2=0.212) while negative feedback does not affect it by much (and in fact may increase it) (ω_4=0.522).

The second row in Table 2 illustrates the results for how the different experiences affect the participants' self-confidence. When receiving feedback on their own move, there are strong positive and negative impact on self-confidence (ω_2=0.828 and ω_4=0.286, respectively) as expected. However, when the participants receive feedback on the AI instead, feedback does not translate to self-confidence in an expected

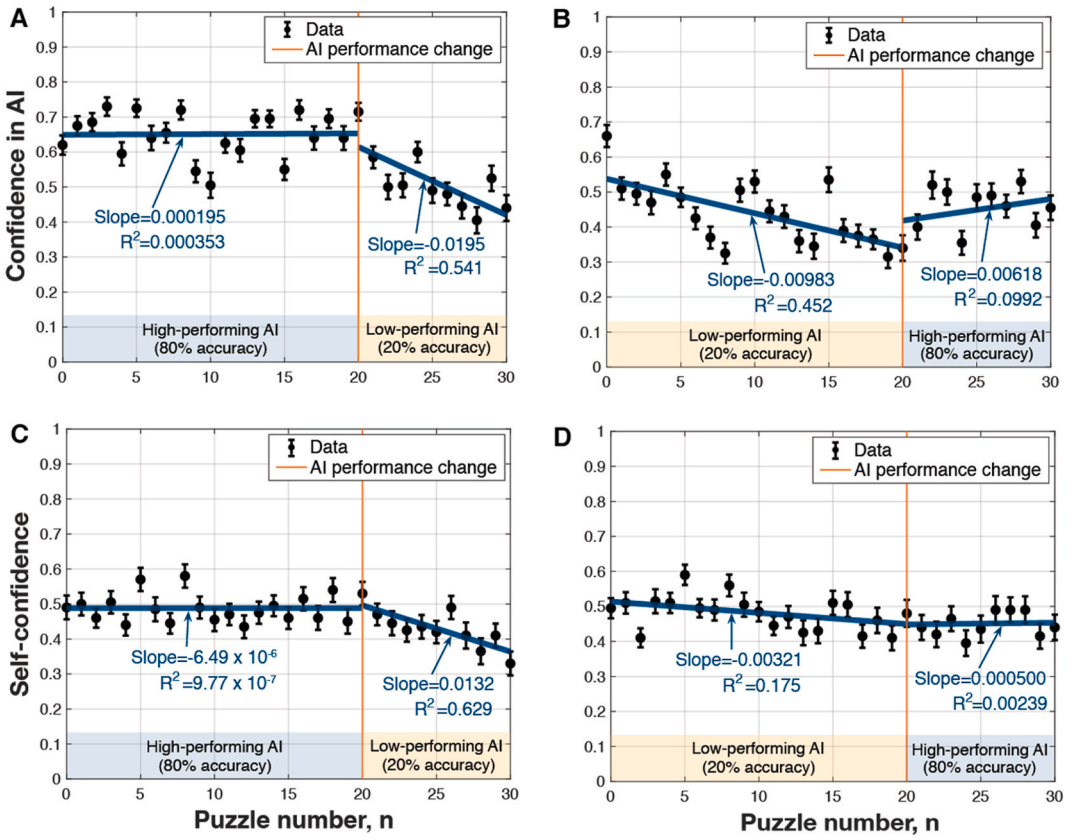

Fig. 4. Confidence in AI and self-confidence plots of both experimental conditions. Similar to Fig. 3, (A) and (B) are human confidence in the AI plots for Conditions 1 and 2 respectively, and (C) and (D) are human self-confidence plots for Conditions 1 and 2 respectively. Black data points are the average confidence values of the 50 participants in each trial. The dark blue lines are the linear fit to the average confidence data before and after the AI performance change (orange line). These dark blue lines represent the general trend of incline or decline of the respective confidence. The error bars represent the standard error of the data.

Table 2
Model parameter estimates corresponding to the impact of the four types of experiences in AI-assisted decision-making on the participants' confidence in the AI and their self-confidence. The quantitative model of confidence in AI and self-confidence are fitted to the experimental data. Parameter values ω_1 and ω_2 correspond to experiences of receiving positive feedback on the AI suggestion and on their own move (i.e., $e_1(n)$ and $e_2(n)$ in Fig. 2), respectively. Similarly, parameter values ω_3 and ω_4 correspond to those of receiving negative feedback on the AI suggestion and on their own move (i.e., $e_3(n)$ and $e_4(n)$ in Fig. 2), respectively. Table 1 shows the complete parameter fitting results.

	Impact of the four types of experiences			
	Positive feedback		Negative feedback	
	AI (ω_1)	Self (ω_2)	AI (ω_3)	Self (ω_4)
Confidence in AI	0.844	0.212[a]	~0	0.522
Self-confidence	0.574	0.828	0.238[a]	0.286

[a] Major findings that are elaborated in Discussion.

manner. Positive feedback on the AI has low impact on human self-confidence (ω_1=0.574) while negative feedback greatly decreases it (ω_3=0.238). These results mean that if the participants accept good AI suggestions, they are slightly more confident in themselves, and if they accept bad ones, they lose a significant degree of confidence in themselves.

3.3. Impact of human confidence in AI and human self-confidence on their AI acceptance decisions

The binary data on the decision to accept or reject the AI suggestions are logistically regressed against the confidence in the AI and self-confidence data to analyze the impact of these two types of human confidence on their AI acceptance decisions. The logistic regression includes two predictor variables (confidence in the AI and self-confidence), 99 dummy variables (100 subjects), and one response variable (decision to accept or reject AI suggestion). Results show that the participants' confidence in the AI does not significantly affect whether they accept or reject the AI suggestion (coefficient=0.150, P=0.5). Rather, it is their self-confidence that determines this decision (coefficient=−1.00, P<0.001). These results are very surprising because they overturn the intuitive assumption that humans accept AI suggestions when they are confident in the AI's ability and vice versa; instead, our results show that human self-confidence, not confidence in the AI, directs their decisions to accept or reject AI suggestions.

3.4. Characteristics of successful decision-makers

Considering teams each consisting of a human decision-maker and an AI advisor, the decision-making patterns of the participants with low, mid, and high final team performance are compared. The final team performance score represents how successfully the participants have made the final move. The quality of the final move shows whether the participants appropriately accepted or rejected the AI suggestions; therefore, low-, mid-, and high-performance teams correspond to poor, fair, and good human decision-makers. For each experimental condition, the final team performance scores of its participants are fit to a normal distribution (see Fig. 5). Then, each condition's distribution is divided into three sections: low 25%, high 25%, and the rest in the middle, shown as the orange lines in Fig. 5. For Condition 1, there are 8, 10 and 32 participants in low, high, and middle sections respectively, while in Condition 2, there are 7, 11, and 32 participants in the corresponding

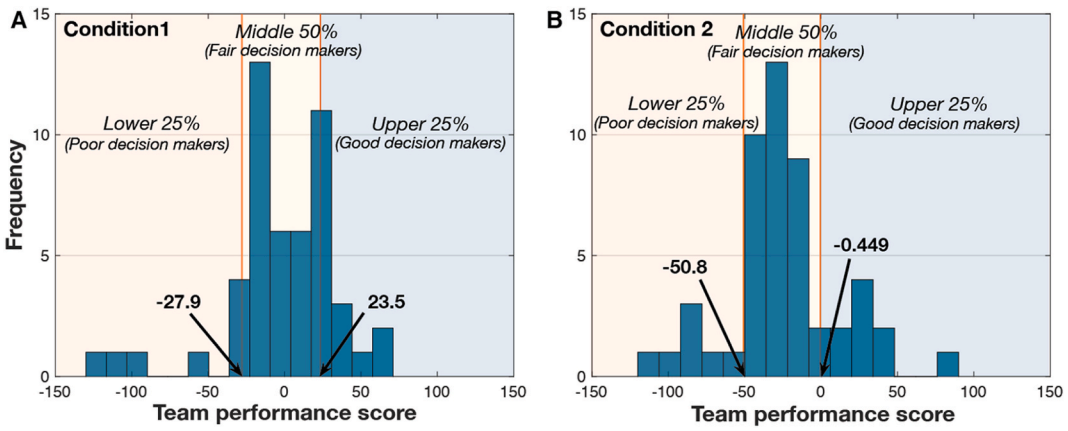

Fig. 5. Histograms of team performance score distribution among the participants in each experimental condition. (A) and (B) show the histograms for Condition 1 and 2, respectively. The orange lines are the boundaries for the lower 25% cut and the upper 25% cut. Participants whose performance is located to the left of the lower orange line, in between the orange lines, and to the right of the upper orange line are classified as poor, fair, and good decision-makers, respectively.

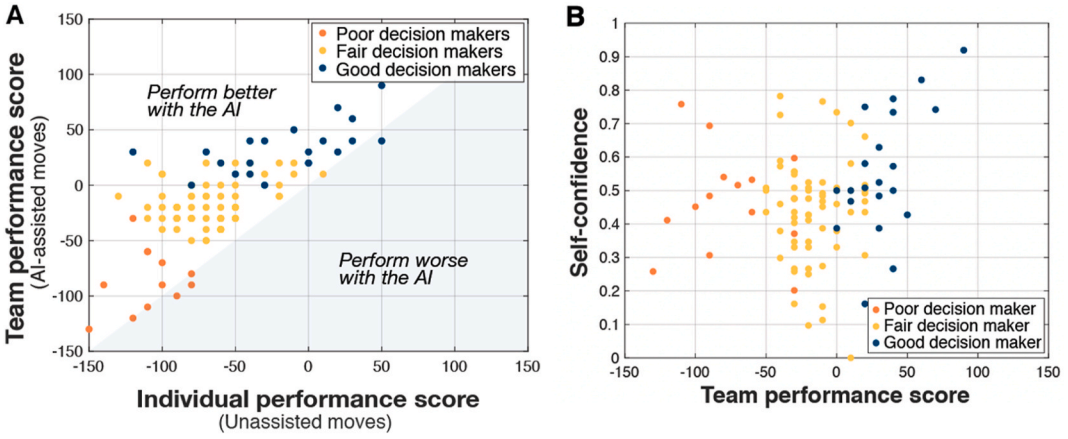

Fig. 6. Comparison of characteristics among poor, fair, and good decision-makers. (A) Individual vs. team performance score of the participants. Each data point on the plot corresponds to one participant. Individual performance scores are calculated using the participants' unassisted selections before they received AI suggestions. Team performance scores are calculated using the participants' moves after receiving AI suggestions. (**B**) Self-confidence vs. team performance score plot. The plot shows the average self-confidence of the 100 participants. Each data point corresponds to each participant's average self-confidence in the experiment.

Table 3
Regression results between human confidence and AI acceptance decisions for poor, fair, and good decision-makers. Results in this table are from the combined data from the two experimental conditions. The first row of the table shows the regression coefficients between the participants' confidence in the AI and their probability of accepting the AI suggestions. The second row shows the regression coefficients between the participants' self-confidence and their probability of accepting the AI suggestions.

	Regression coefficient against the probability of accepting AI suggestion		
	Low (Poor)	Mid (Fair)	High (Good)
Confidence in AI	−0.0737 *(P=0.9)*	0.0820 *(P=0.7)*	−0.691 *(P=0.2)*
Self-confidence	−1.36 *(P=0.09)*	−0.813 *(P<0.05)*	1.78 *(P<0.05)*

sections. Finally, the participants who fall into each section in the two experimental conditions are referred to as poor, good, and fair decision-makers; therefore, there are total 15, 21, and 64 participants in each category respectively. The data from the two conditions are combined to construct the results in Fig. 6 and Table 3.

Note that not all the poor decision makers' team performance scores are lower than those of the good decision makers because these classifications are done separately for the two conditions of the experiment. In the first condition (Fig. 5A), poor decision makers all have team performance scores lower than −27.9, while those in the second condition (Fig. 5B) have scores lower than −50.8. Additionally, Fig. 6A and B below show differing number of data points plotted for each color group because of the overlapping data points. For example, in Fig. 6A, there is a pair of participants with the same team and individual performance scores among the poor decision makers, resulting in the two data points overlapping into one. These participants however do not have the same self-confidence, therefore showing as two separate data points in Fig. 6B.

Fig. 6A shows the relationship between final team performance and individual skill level. Poor decision-makers (i.e., those with low team performance scores) are relatively poor chess players, while good decision-makers are spread throughout poor to good chess players. This may lead to an assumption that poor and good decision-makers have different ranges of self-confidence in their chess skills. However, the results in Fig. 6B demonstrate that regardless of their skill level, the participants' self-confidence vary over a similar range, meaning good decision-makers can have low self-confidence while poor decision-makers can have high self-confidence.

In addition to the earlier finding that human self-confidence directs their decision to accept or reject the AI suggestion, poor, fair, and good decision-makers do not show clear differences in their overall self-

confidence. This suggests that the three groups perform differently due to the differences in how they translate their self-confidence levels to decisions, not their self-confidence itself. Therefore, regression results between the participants' self-confidence and AI acceptance decisions are compared among the poor, fair, and good decision-makers. While all these three groups show no significant relationship between their confidence in the AI and AI acceptance decisions, there are clear variations in the self-confidence results between the groups (Table 3). Notably, good decision-makers uniquely exhibit a positive relationship between their self-confidence and the probability of accepting AI suggestions, while the other participants show a negative relationship. This means that good decision-makers differ from poorer decision makers in that they show a decision pattern of accepting the AI suggestions when they are confident in themselves and rejecting them when they are not.

4. Discussion

This paper began by asking how changing AI performance affect human confidence in AI and human self-confidence, given the notion that they are prone to changes based on the performance of the AI (Hancock et al., 2011; Hu et al., 2019; Schoorman et al., 2007). The results reveal the detrimental effect of poor AI performance on both human confidence in the AI and their self-confidence. First, although the participants initially have relatively high confidence in the AI, poor AI performance quickly decreases this confidence. This is a significant problem because the confidence in the AI is gained back slowly (even with subsequent high AI performance) but is rapidly lost. This result confirms prior studies that demonstrated that humans tend to show "high" initial trust in embedded AI which is difficult to increase but decreases with AI error (de Visser et al., 2017; Dietvorst et al., 2015; Glikson & Woolley, 2020). This asymmetric impact of low and high AI performance on confidence in the AI may be explained by the notion of loss aversion in prospect theory: losses loom larger than corresponding gains (Tversky & Kahneman, 1991). Although the participants are not explicitly informed of the AI performance change, accepting suggestions from a poor-performing AI is often followed by a loss of points. As a result, this "loss" impacts their confidence in the AI more than the corresponding "gain" with a high-performing AI. Second, poor AI performance also has a negative influence on human self-confidence. When the AI changes to perform poorly, humans lose confidence in the AI *and in themselves.* While it is appropriate to penalize the AI for its poor performance, humans also penalize themselves perhaps for failing to detect the AI error and accepting its suggestions.

Our work also shows that humans tend to *misattribute credit and blame when they infer information* from their experience, as is the case in this study. For example, when the experience provides direct information about the AI, humans also *infer* what this experience tells them about themselves; similarly, when the experience is about themselves, they also *infer* insight into the AI. Misattribution is first observed when humans reject the AI suggestion and receive positive feedback on their own performance. Although this positive feedback is on their own performance, the participants lose confidence in the AI. This may be because when humans are affirmed in their ability, their increased self-confidence leads them to look down on the AI. In addition, when they accept the AI suggestion and receive negative feedback, humans lose their self-confidence, by which it can be inferred that they are attributing blame to themselves. This is consistent with the earlier discussion that humans penalize not only the AI but also themselves for negative AI performance. In response to negative feedback on the AI, they may be blaming themselves for failing to detect the AI error. While this misattribution of blame can be beneficial in a managerial perspective as humans are properly taking responsibility as the final decision-maker, it could be a factor leading humans to inappropriately accept or reject AI suggestions. The answers to the first research question open doors for strategic maneuvering of AI and feedback to appropriately calibrate human confidence, such as decreasing AI accuracy and providing positive feedback on human moves to lower their unfittingly high confidence in an AI.

Understanding human confidence dynamics during AI-assisted decision-making leads us to the second research question: how are human confidence in the AI and their self-confidence associated with the probability of accepting AI suggestions? The result from our study surprisingly concludes that *human self-confidence significantly contributes to their acceptance of AI decisions, while their confidence in the AI does not.* Although self-confidence has repeatedly been recognized as an important factor affecting the willingness to rely and therefore human-AI trust and use (Bagheri & Jamieson, 2004; Dzindolet et al., 2002; Lee & Moray, 1994), this work uniquely identifies self-confidence as a more powerful factor than confidence in the AI. For instance, while Lee and Moray have shown the logit relationship between the difference between trust and self-confidence, and the use of automation (Lee & Moray, 1994), they do not inform about the individual influence of trust and self-confidence on the use of automation. However, our work provides insight that self-confidence is most likely the driving source of the logit relationship. Therefore, the significant correlation found between human self-confidence and the decisions to accept or reject AI suggestions highlights the importance of managing human self-confidence over their confidence in the AI for successful AI-assisted decision-making. By skillfully orchestrating AI performance and human experience using the insights from the earlier discussion (e.g., recover from human misattribution of blame on themselves by subsequently providing positive feedback on their own moves), self-confidence can be calibrated effectively.

Finally, this work identifies that *good decision-makers uniquely display a positive correlation between self-confidence and probability of accepting AI suggestions*; they accept the AI when they are confident in themselves and reject the AI when they are not. Although good decision-makers show similar self-confidence levels as any others, their decision pattern successfully translates self-confidence to decisions. Earlier discussion shows that when humans mistakenly accept a poor AI suggestion and receive negative feedback, they lose self-confidence (i.e., human misattribution). Then, poor and fair decision-makers' probability of accepting the next AI suggestion increases, causing them to enter a vicious cycle of relying on a poorly performing AI. In contrast, good decision-makers' probability of accepting the next AI suggestion decreases, avoiding the vicious cycle. Therefore, methods may be developed to identify good decision-makers with such positive correlation between self-confidence and probability of accepting AI suggestions, or to cultivate decision-making strategies to prevent the vicious cycle of relying on a poorly performing AI.

This work has some limitations that offer opportunities for future research. First, the study focuses specifically on the chess puzzles task. Extending this study to different human-AI problem-solving applications will indicate whether the dynamics of human confidence and its effect on adoption of AI advice differ in higher-stakes situations. Additionally, the results could be relevant only to this specific experimental design where the participants are explicitly and frequently reporting their confidence levels. Such explicit repeated reporting does not perfectly resemble real-world situations and may influence people's subsequent judgements (Kvam et al., 2015). Therefore, it would be beneficial to extend this study to use more inconspicuous approaches to measure confidence in AI and self-confidence. Finally, this work focuses on two types of human confidence that affect trust: confidence in AI and self-confidence. Although this focus is beneficial in gaining detailed insights, for a comprehensive understanding of how trust affects AI-assisted decision-making, other factors of trust such as personality and environment can be explored.

5. Conclusion

Overall, the results of this work indicate a human tendency to misattribute the blame for poor AI performance to themselves, a

significant impact of human self-confidence on their decisions to accept or reject AI suggestions, and a resulting vicious cycle that hinders effective human-AI decision-making. This research shows that poor AI performance decreases human self-confidence which is found to influence the decision to accept or reject AI suggestions. This misattribution exposes many human decision-makers to a vicious cycle of relying on a poorly performing AI. Although good decision-makers can break out of this cycle, many others cannot as their decreased self-confidence from the misattribution inclines them to accept the next suggestion from a poorly performing AI. Our results directly affect the success of AI-assisted decision-making by providing insight into the cause of human mis-reliance on AI. These results also inspire new strategies for confidence calibration to reduce such mis-reliance. Finally, they shine light on the significance of human self-confidence in AI-assisted decision-making.

Credit author statement

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

Declaration of competing interest

None.

Acknowledgments

This work was supported by the Air Force Office of Scientific Research (AFOSR) under grant FA9550-18-1-0088. The sponsor had no other involvement beyond the financial support.

References

Alvarado-Valencia, J. A., & Barrero, L. H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior, 36*, 102–113. https://doi.org/10.1016/j.chb.2014.03.047

Bagheri, N., & Jamieson, G. A. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced 'complacency'. *Human Performance, Situation Awareness, and Automation Current Research Trends, 2*, 54–59.

Bansal, G., et al. (2019a). Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of AAAI conference on human computation and crowdsourcing* (Vol. 7, pp. 2–11).

Bansal, G., et al. (2019b). Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of AAAI conference on artificial intelligence* (Vol. 33, pp. 2429–2437). https://doi.org/10.1609/aaai.v33i01.33012429

Buch, V. H., Ahmed, I., & Maruthappu, M. (2018). Artificial intelligence in medicine: Current trends and future possibilities. *British Journal of General Practice, 68*, 143–144. https://doi.org/10.3399/bjgp18X695213

Crisp, C. B., & Jarvenpaa, S. L. (2013). Swift trust in global virtual teams. *Journal of Personnel Psychology, 12*, 45–56. https://doi.org/10.1027/1866-5888/a000075

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In *Proceedings of the 11th international joint Conference on artificial intelligence* (pp. 4691–4697). IJCAI. https://doi.org/10.24963/ijcai.2017/654.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*, 114–126. https://doi.org/10.1037/xge0000033

Dietz, G., & Den Hartog, D. N. (2006). Measuring trust inside organisations. *Personnel Review, 35*, 557–588. https://doi.org/10.1108/00483480610682299

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

Dzindolet, M. T., et al. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*, 79–94. https://doi.org/10.1518/0018720024494856

Dzindolet, M. T., et al. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*, 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

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

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*, 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Hancock, P. A., et al. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*, 517–527. https://doi.org/10.1177/0018720811417254

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems, 28*, 84–88. https://doi.org/10.1109/MIS.2013.24

Hu, W. L., et al. (2019). Computational modeling of the dynamics of human trust during human-machine interactions. *IEEE Transactions on Human-Machine Systems, 49*, 485–497. https://doi.org/10.1109/THMS.2018.2874188

Jonker, C. M., & Treur, J. (1999). Formal analysis of models for the dynamics of trust based on experiences. In *Proceedings of 9th European Workshop on modelling autonomous Agents in a multi-agent world* (pp. 221–231). MAAMAW. https://doi.org/10.1007/3-540-48437-X_18.

Kamar, E., Hacker, S., & Horvitz, E. (2012). Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th international Conference on autonomous Agents and multiagent systems* (pp. 467–474). AAMAS.

Kvam, P. D., et al. (2015). Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. In *Proceedings of the national academy of sciences* (Vol. 112, pp. 10645–10650). https://doi.org/10.1073/pnas.1500688112

Lee, J. D., & Moray, N. (1994). Trust, self-Confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*, 153–184. https://doi.org/10.1006/ijhc.1994.1007

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

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management, 32*, 991–1022. https://doi.org/10.1177/0149206306294405

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social Forces, 63*, 967–985. https://doi.org/10.1093/sf/63.4.967

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*, 709–734. https://doi.org/10.2307/258792

Mcknight, D. H., & Chervany, N. L. (1996). *The meanings of trust.* Techical Report MISRC Working Paper.

Moré, J. J., & Sorensen, D. C. (1983). Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing, 4*, 553–572. https://doi.org/10.1137/0904038

Nagar, Y., & Malone, T. W. (2011). Making business predictions by combining human and machine intelligence in prediction markets. In *Proceedings of international conference on information systems* (pp. 4–7). ICIS.

Parasuraman, R., Cosenzo, K. A., & De Visser, E. (2009). Adaptive automation for human supervision of multiple uninhabited vehicles: Effects on change detection, situation awareness, and mental workload. *Military Psychology, 21*, 270–297. https://doi.org/10.1080/08995600902768800

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency". *The International Journal of Aviation Psychology, 3*, 1–23. https://doi.org/10.1207/s15327108ijap0301_1

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*, 230–253. https://doi.org/10.1518/001872097778543886

Patel, B. N., et al. (2019). Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *Npj Digital Medicine, 2*, 111. https://doi.org/10.1038/s41746-019-0189-7

Richtel, M., & Dougherty, C. (2015). *Google's driverless cars run into problem: Cars with drivers*. New York Times. https://www.nytimes.com/2015/09/02/technology/personaltech/google-says-its-not-the-driverless-cars-fault-its-other-drivers.html.

Rousseau, D. M., et al. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*, 393–404. https://doi.org/10.5465/AMR.1998.926617

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of Management Review, 32*, 344–354. https://doi.org/10.5465/AMR.2007.24348410

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal, 31*, 47–53.

Strickland, E. (2019). *How IBM Watson overpromised and underdelivered on AI health care*. IEEE Spectrum. https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics, 106*, 1039–1061. https://doi.org/10.2307/2937956

de Visser, E. J., 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*, 116–133. https://doi.org/10.1177/0018720816687205

Wilson, H. J., & Daugherty, P. R. (2018). *Collaborative intelligence: Humans and AI are joining forces*. Harvard Business Review. https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces.

Yin, M., Vaughan, J. W., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of ACM conference on human factors in computing systems* (pp. 1–12). CHI. https://doi.org/10.1145/3290605.3300509.

Zhang, G., et al. (2021). A cautionary tale about the impact of AI on human design teams. *Design Studies, 72*, 100990. https://doi.org/10.1016/j.destud.2021.100990

Zhang, Y., Vera Liao, Q., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of ACM conference on fairness, accountability, and transparency* (pp. 295–305). FAT*. https://doi.org/10.1145/3351095.3372852.