Leah Chong
Mem. ASME
Department of Mechanical Engineering,
Carnegie Mellon University,
5000 Forbes Ave.,
Pittsburgh, PA 15213
e-mail: lmchong@andrew.cmu.edu

Ayush Raina
Mem. ASME
Department of Mechanical Engineering,
Carnegie Mellon University,
5000 Forbes Ave.,
Pittsburgh, PA 15213
e-mail: araina@andrew.cmu.edu

Kosa Goucher-Lambert
Mem. ASME
Department of Mechanical Engineering,
University of California,
6179 Etcheverry Hall,
Berkeley, CA 94720
e-mail: kosa@berkeley.edu

Kenneth Kotovsky
Department of Psychology,
Carnegie Mellon University,
5000 Forbes Ave.,
Pittsburgh, PA 15213
e-mail: kotovsky@cmu.edu

Jonathan Cagan[1]
Department of Mechanical Engineering,
Carnegie Mellon University,
5000 Forbes Ave.,
Pittsburgh, PA 15213
e-mail: cagan@cmu.edu

The Evolution and Impact of Human Confidence in Artificial Intelligence and in Themselves on AI-Assisted Decision-Making in Design

Decision-making assistance by artificial intelligence (AI) during design is only effective when human designers properly utilize the AI input. However, designers often misjudge the AI's and/or their own ability, leading to erroneous reliance on AI and therefore bad designs occur. To avoid such outcomes, it is crucial to understand the evolution of designers' confidence in both their AI teammate(s) and themselves during AI-assisted decision-making. Therefore, this work conducts a cognitive study to explore how to experience various and changing (without notice) AI performance levels and feedback affects these confidences and consequently the decisions to accept or reject AI suggestions. The results first reveal that designers' confidence in an AI agent changes with poor, but not with good, AI performance in this work. Interestingly, designers' self-confidence initially remains unaffected by AI accuracy; however, when the accuracy changes, self-confidence decreases regardless of the direction of the change. Moreover, this work finds that designers tend to infer flawed information from feedback, resulting in inappropriate levels of confidence in both the AI and themselves. Confidence in AI and self-confidence are also shown to affect designers' probability of accepting AI input in opposite directions in this study. Finally, results that are uniquely applicable to design are identified by comparing the findings from this work to those from a similar study conducted with a non-design task. Overall, this work offers valuable insights that may enable the detection of designers' dynamic confidence and their consequent misuse of AI input in the design.

[DOI: 10.1115/1.4055123]

Keywords: artificial intelligence, cognitive-based design, collaborative design, design methodology

1 Introduction

Artificial intelligence (AI) has extended its influence to design settings by demonstrating its potential to contribute to various steps of the design process such as customer preference identification, ideation, and manufacturing [1–4]. One example is the data-driven 3D shape generation algorithm presented by Zhang et al. which creates a variety of new designs from a given dataset [5]. Furthermore, Raina et al. developed an AI that learns to design from human data and generates products that are as good as those created by human designers [6]. Some AI can even outperform human designers [7,8]. However, AI systems possess their own set of limitations that may be complemented by human strengths, suggesting an opportunity for human-AI collaboration in design. For example, AIs have an advantage in their ability to efficiently pull insights from large data but cannot yet replace human agility and creativity [9]. Together, human designers and AI may be able to solve complex, dynamic problems that neither of them can solve alone [10,11].

An important question about human-AI collaboration in design is how to design and enable an AI agent to aid human designers as effective and empowering teammates. One way may be by assisting the human decision-making; such human-AI teaming is prevalently known as AI-assisted decision-making. During AI-assisted decision-making, humans receive design suggestions from their AI teammates, which they must either accept or override to make the final decision. Here, AI is contributing to the team as a "second opinion" system, providing its best solution to humans as a second opinion to consider.

AI assistance, however, is only beneficial when human designers appropriately utilize AI input. Unfortunately, humans often fail to discern when to accept or reject AI suggestion(s), defeating the purpose of AI-assisted decision-making. Further, such errors can negatively impact human lives and, more likely, the quality of human lives. In non-design, high-stakes contexts, faulty judgment of AI input has already been shown to have detrimental consequences [12–14]. For example, in 2015, when a Google self-driving car was slowing down for a pedestrian, the driver further applied the brakes, causing the car to be hit from behind [15]. Such an outcome is very likely also in design decision-making scenarios. For example, when designing seats for an airplane, accepting a poor design suggestion from an AI can significantly aggravate user experience and might even contribute to the loss of human lives in case of an accident.

Prior research has shown that human designers' decision to accept or reject AI suggestion(s) is dependent on their trust in the AI [12–14,16–20], meaning that humans make erroneous decisions when their trust does not match the AI's trustworthiness. Many factors influence this trust, especially those that affect their perception of the AI or themselves. As in our earlier work [21], the current

[1]Corresponding author.

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received January 29, 2022; final manuscript received July 11, 2022; published online October 31, 2022. Assoc. Editor: Daniel A. McAdams.

work investigates two of these factors in the context of design decision-making: confidence in AI and self-confidence. Confidence in AI and self-confidence bring insight into two antecedents of trust proposed by Mayer et al.: perception of trustee's attributes such as ability, and a propensity to trust [22,23]. Therefore, confidence in AI represents humans' perception of AI's task ability, while self-confidence represents the perception of their own task ability, contributing to humans' inclination to rely on the AI.

Chong et al. [21] studied human confidence in AI and self-confidence during an AI-assisted chess puzzle task and provided valuable insight into the evolution of human confidence and its impact on decision-making. However, the results from their work are specific to tasks sharing similar characteristics as the chess puzzle task and therefore may not apply directly to design tasks. Furthermore, the two types of human confidence have been understudied in design contexts likely due to the subjectivity in defining what "good" design is and consequently how good the AI suggestions are. An accurate understanding of confidence dynamics and its influence on the decision to accept AI input during design tasks is critical to resolving human designers' erroneous reliance on AI and improving the effectiveness of AI-assisted decision-making in design.

To achieve this understanding, this work investigates the evolution of human confidence in AI and human self-confidence, and their impact on AI-assisted decision-making during a truss design task. Insights into human designers' confidence during AI-assisted decision-making can inform effective formation of human-AI design teams and improve the overall outcome of teamwork by suggesting ways to reduce inappropriate reliance on AI. Although the truss design task may not represent all types of design problems, it is chosen for its well-defined and sequential nature. In contrast to some other design problems, it is possible to define and evaluate good design and good AI suggestions in the truss design problem, which is crucial in supplying objective feedback to the participants. Moreover, the sequential nature of the truss design task not only provides a great setting to study dynamic confidence but also resembles the chess puzzle task and enables comparison between the results of the two studies. The current work explores the same research questions from Chong et al. [21] but in the context of design as follows:

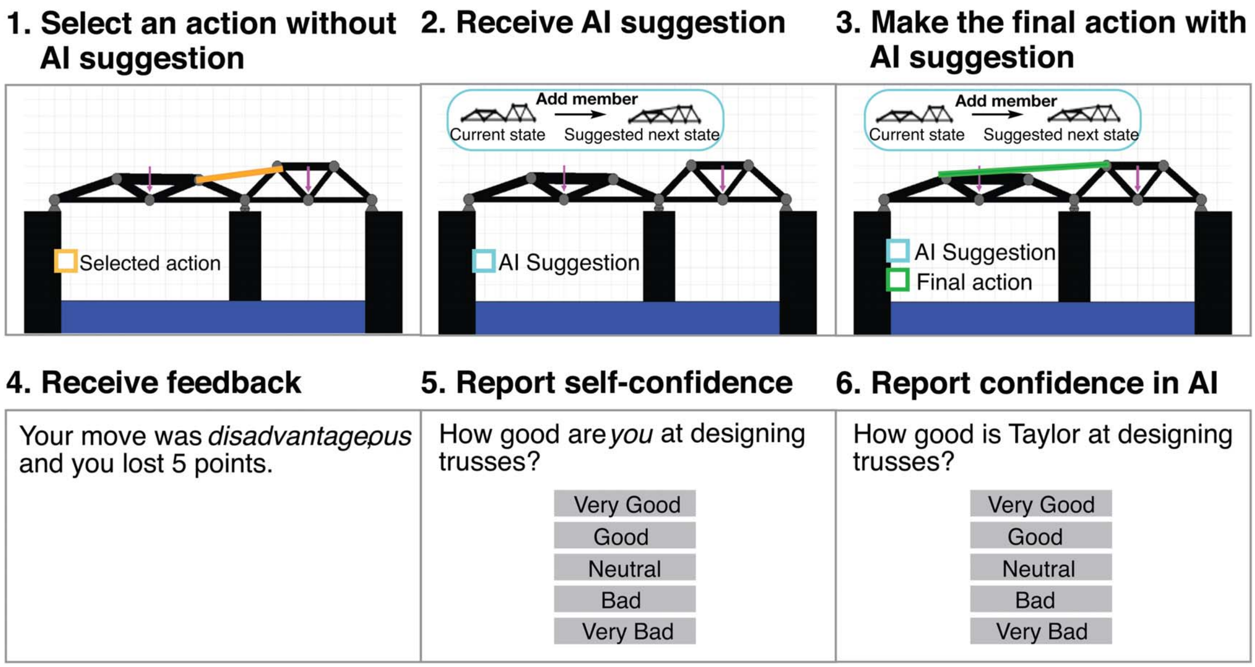
(1) How to do changes in AI performance and resulting positive and negative feedbacks affect human confidence in AI and human self-confidence?
(2) How are these two types of confidence associated with the probability of accepting AI suggestions?
(3) What decision-making patterns distinguish those who successfully accept and reject AI suggestions? and additionally answers the following question:
(4) Which of the results are unique to design?

2 Methods

For the purpose of this work, a human subject study and a quantitative model are used to collect and model the data which are then analyzed via statistical methods. The cognitive study shows how human confidence in AI and human self-confidence evolves over the course of AI-assisted decision-making in design in view of changes in AI accuracy. The quantitative model of human confidence captures the impact of various experiences during AI-assisted decision-making on the two types of confidence.

2.1 Human Subject Study

2.1.1 Experimental Task. Participants are given a truss design task in which they must make the best next action given a truss state (see Fig. 1 for an example problem). They are asked to work with an AI teammate, named Taylor, who provides an action suggestion for each problem. In this task, the "best" action means the most advantageous action at the given moment to maximize the strength-to-weight ratio (SWR) of the truss. This truss design task used in this study originates from earlier work by McComb et al. where the goal was to create optimal trusses that can handle certain loads [24]. The original task is modified by eliminating the "add node" and "move node" options to reduce the design space, limiting the search of the AI algorithm and making it possible to evaluate the goodness scores of all possible actions. There are always finite numbers of possible actions because the only action options are "delete node", "add member", "delete member", "decrease member thickness", and "increase member thickness", the latter two at defined increments. There is total of 33 truss



Your move was *disadvantagepus* and you lost 5 points.

How good are *you* at designing trusses?

Very Good
Good
Neutral
Bad
Very Bad

How good is Taylor at designing trusses?

Very Good
Good
Neutral
Bad
Very Bad

Fig. 1 Truss design task procedure. The user interface is adopted from McComb et al. [24] and modified for the needs of this study.

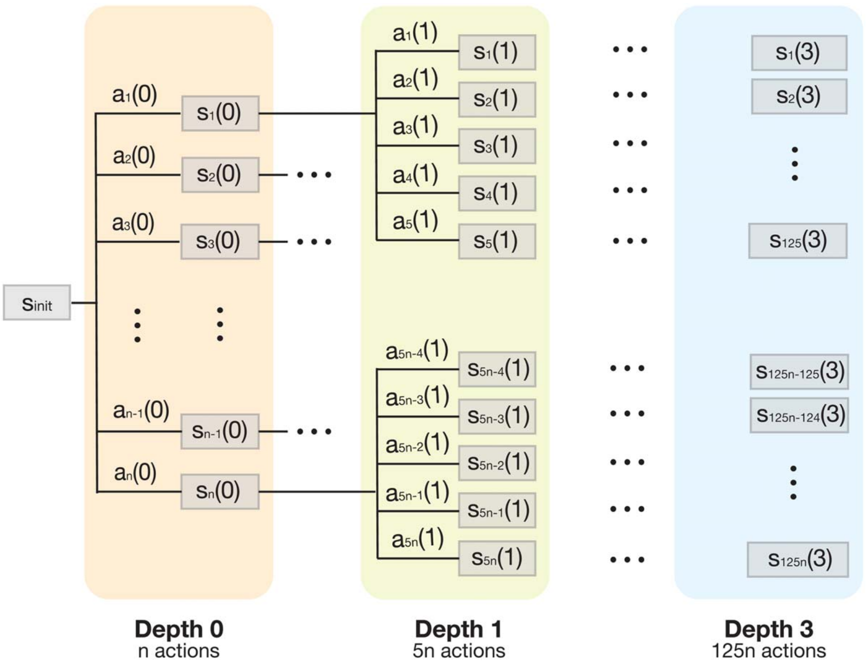


Fig. 2 Tree search AI algorithm that evaluates the goodness score of each action. For each truss problem (or each initial truss state, s_{init}), the AI conducts a tree search to Depth 3 to evaluate the goodness of each possible action from that state. Variables *s* and *a* respectively represent state and action, and the depths are indicated in parentheses.

design problems in this study (three for practice and 30 for the experiment) to capture AI-assisted decision-making scenarios with sequence of design actions and examine how the experience in each problem influences human confidence in AI and in themselves. It is ensured that all 30 problems have at least one good action and one bad action, distinguishable by the following method to calculate the goodness scores of the actions.

An AI algorithm based on deep learning and tree search is developed to determine good and bad actions from a given state and provide suggestions and feedback to the participants (see Fig. 2). In the first step, given a truss state (s_{init}), all possible discrete actions are determined (see Depth 0 in Fig. 2). Each of these actions results in a new truss state. To determine the goodness score of these actions, their resulting states are evaluated using a tree search. The computational cost associated with the tree search is $>10^6$ states for Depth 3, which is effectively higher than what humans can search. To mitigate the cost, rather than evaluating all possible actions, the tree search leverages a data-driven agent from Raina et al. [6] to select five meaningful actions at each depth. The agent generates a list of all possible actions at a given state and ranks them in the order of likeliness that a human will make the action. Our AI algorithm utilizes this likeliness ranking to select the top five meaningful actions at each depth and achieve a search to Depth 3. The goodness score of all possible actions in Depth 0 are calculated using the resulting states at Depth 3. For each action in Depth 0, its goodness score is determined by calculating the weighted sum of the SWR of 5^3 or 125 truss states at Depth 3 that branched off from the action.

2.1.2 Participants. 100 participants were recruited for (and completed) the experiment following a protocol approved by Carnegie Mellon University's Institutional Review Board. All participants are Mechanical Engineering undergraduates or graduate students from Carnegie Mellon University or University of California at Berkeley. Prior to their participation, they have all taken a mechanics course, and therefore they had experience with designing trusses. It is worth noting that although there may be significant differences in the truss design skills among undergraduates and graduate students, this is not a problem for the purpose of the study because most analyses are based on the averaged data over a large pool. Moreover, rather than controlling their truss design skills, the participants' individual skill levels are collected during the experiment.

2.1.3 Experimental Conditions. There is total of 30 truss design problems in the study. After problem 20, the quality of AI suggestions changes to instantiate the dynamic performance of the AI, and the direction of this change distinguishes the two experimental conditions. Depending on the similarity of task scenarios to AI's training dataset, the AI performance may fluctuate. In the first condition (hi-lo—meaning the AI begins with primarily high accuracy, and transitions to low), AI accuracy switches from 80% to 20%, while in the second condition (lo-hi—meaning the AI begins with primarily low accuracy, and transitions to high), it changes from 20% to 80%. 80% accuracy means that 80% of the time, the AI chooses the best action out of all possible actions. For the other 20% of the times, the AI chooses the worst action. In the same way, 20% accuracy means that the AI chooses the best action 20% of the time. Each participant is randomly assigned to one of the two experimental conditions (50 per condition).

2.1.4 Procedure. The experiment is conducted virtually on Amazon Web Services. Prior to participation, informed consent is obtained via Google Forms from all participants. The participants then receive an email with a step-by-step instruction of the experiment and are assigned 90 min to complete the experiment.

All participants follow the same procedure for each truss design problem, as shown in Fig. 1. First, given a truss state, they select their best next action before receiving an AI suggestion (first step in Fig. 1). This unassisted action is designed into the experiment to collect data about the participants' independent truss design task ability. After the unassisted action is selected, the participants receive an AI suggestion (second step in Fig. 1). Considering this suggestion, they make the final decision to either accept it or override it with another action (third step in Fig. 1). When overriding, the participants are not limited to their initial, unassisted action but can make any action different from the AI suggestion. Next, the participants gain either positive or negative feedback depending on whether the final action is advantageous or disadvantageous towards a high SWR (fourth step in Fig. 1). This advantage and disadvantage are determined by the goodness evaluation score introduced earlier (see Fig. 2). According to the feedback, the participants also gain or lose five points. At the start of the experiment, to incentivize the participants, they are informed that those who receive a score higher than ten will receive an additional monetary prize. This threshold is a score that was difficult to be reached by participants in the pilot study. Finally, for each problem, the participants report their confidence in their own ability and in the AI's ability to design trusses, each in a five-point Likert scale (last two steps in Fig. 2). The confidence questions ask: How good are you (or the AI) at designing trusses? and the answers include very good, good, neutral, bad, and very bad, which are respectively quantified as 1, 0.75, 0.5, 0.25, and 0.

Once the participants are done with the 33 problems (the first three are practice problems), they are asked to fill out a post-experiment questionnaire that is designed to gain more insight into the results of the study. The questionnaire contains five questions which are as follows:

(1) How helpful were the AI suggestions in doing this task?
(2) Was the quality of the AI suggestions consistent? If not, how did it change?
(3) How good were you at designing trusses?
(4) How good were you at making the final decision of which move to choose between your own move and an AI suggestion?
(5) When deciding between your own design move and an AI suggestion, what did you consider more: AI's ability to design trusses or your own ability to design trusses?

2.2 Confidence Model

2.2.1 Model Description. The analysis made in this paper uses the confidence model proposed by Chong et al. [21]. The model's calculation of the trial-by-trial change in human confidence considers three factors: experience, accumulated confidence, and bias, as it did in Hu et al.'s dynamic trust model [25]. Hu et al.'s model was developed and validified to compute the dynamics of human behavioral trust (accept or reject AI inputs) in human-machine interaction contexts. Therefore, the general form of this model is applied to the sequential AI-assisted decision-making context of Chong et al. [21] and this work (see Eq. (1)). However, Chong et al.'s [21] model is unique in its calculation of the experience term (Eq. (2)) because of its application to AI-assisted decision-making scenarios. Hu et al. [25] studied a different, less realistic human-AI decision-making scenario in which humans blindly accept or reject AI suggestions, meaning without much information about the given problem. Furthermore, Chong et al. [21] extended Hu et al.'s [25] model to be applied to self-confidence, in addition to confidence in another person/AI. This section describes information about Chong et al.'s [21] model pertinent to the current work. Additional details can be found in Chong et al. [21].

The following is the general model equation [25]:

$$C(n+1) = C(n) + \alpha_e[E(n) - C(n)] + \alpha_a[A(n) - C(n)] + \alpha_b[B(n) - C(n)] \quad (1)$$

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

The differences between each of the three factors (experience ($E(n)$), accumulated confidence ($A(n)$), and bias ($B(n)$)) and confidence at trial n ($C(n)$) are summed with weights to yield the change in confidence from trial n to $n+1$. α_e, α_a, and α_a are the rate factors.

Experience term in Eq. (1) ($E(n)$) refers to the human designer's experience with the AI at a given trial n. Regardless of the task, there are four types of experiences that can occur in each trial of AI-assisted decision-making as follows:

(1) Accept the AI suggestion, then receive positive feedback (e_1);
(2) Reject the AI suggestion, then receive positive feedback (e_2);
(3) Accept the AI suggestion, then receive negative feedback (e_3);
(4) Reject the AI suggestion, then receive negative feedback (e_4).

In the infrequent case when the participants' unassisted action (before AI suggestion) agrees with the AI suggestion, and they choose this action as the final act, the participants are considered to have accepted the AI suggestion (e_1 or e_3) because they still did not reject it.

In this work, the impact of these four experiences on both human confidence in AI and self-confidence is explored. The experience term at trial n is the weighted (ω_1, ω_2, ω_3, and ω_4) sum of the four sub-experience terms

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

where $e_1(n),\ e_2(n),\ e_3(n),\ e_4(n) = 0$ or 1, $\omega_1, \omega_2, \omega_3, \omega_4 \in [0, 1]$ and $\sum_{n=1}^{4} e_n = 1$.

The exact form of relationship between the sub-experiences is unknown. Therefore, this linear relationship is used to reveal the relative impact of the sub-experiences on the confidence levels when fitted to experimental data. It is important to recognize, however, that the confidence model is overall an iterative, nonlinear model. The aforementioned equations are the model form at each trial n, and the model is fitted to data iteratively over 30 trials.

2.2.2 Parameter Fitting. Parameter estimation was conducted using a trust region reflective algorithm [26] because it is a parameter fitting method for nonlinear models and supports bounded variables. Given initial parameter values, the algorithm iteratively tweaks the parameters and calculates the error until it finds the optimal results. The optimal parameter values fitted to the experimental data from the cognitive study are shown in Table 1. α_e, α_a, α_b are the weights respectively corresponding to the experience, accumulated confidence, and bias terms in the model. ω_1, ω_2, ω_3, ω_4 are the impact factors that represent the impact of the four types of experiences during AI-assisted decision-making on confidence. γ is the time discounting factor in the calculation of the accumulated confidence term. The optimal parameter results were confirmed to be impervious to the initial guess by repeating the estimation process with various initial values.

For a robustness test, the parameters were estimated using 80% of the experimental data (80 participants) that were randomly selected, and this process was repeated 100 times. The initial guesses were set to the values in Table 1. The results from the 100 trials showed only

Table 1 Model parameter fitting results

	α_e	α_a	α_b	ω_1	ω_2	ω_3	ω_4	γ
Confidence in AI	0.329	0.354	0.0433	**0.870**	**0.252**	**0.117**	**0.292**	0.299
Self-confidence	0.286	0.275	0.118	**0.623**	**0.833**	**0.195**	**0.188**	0.279

Note: Bolded results are discussed in this paper.

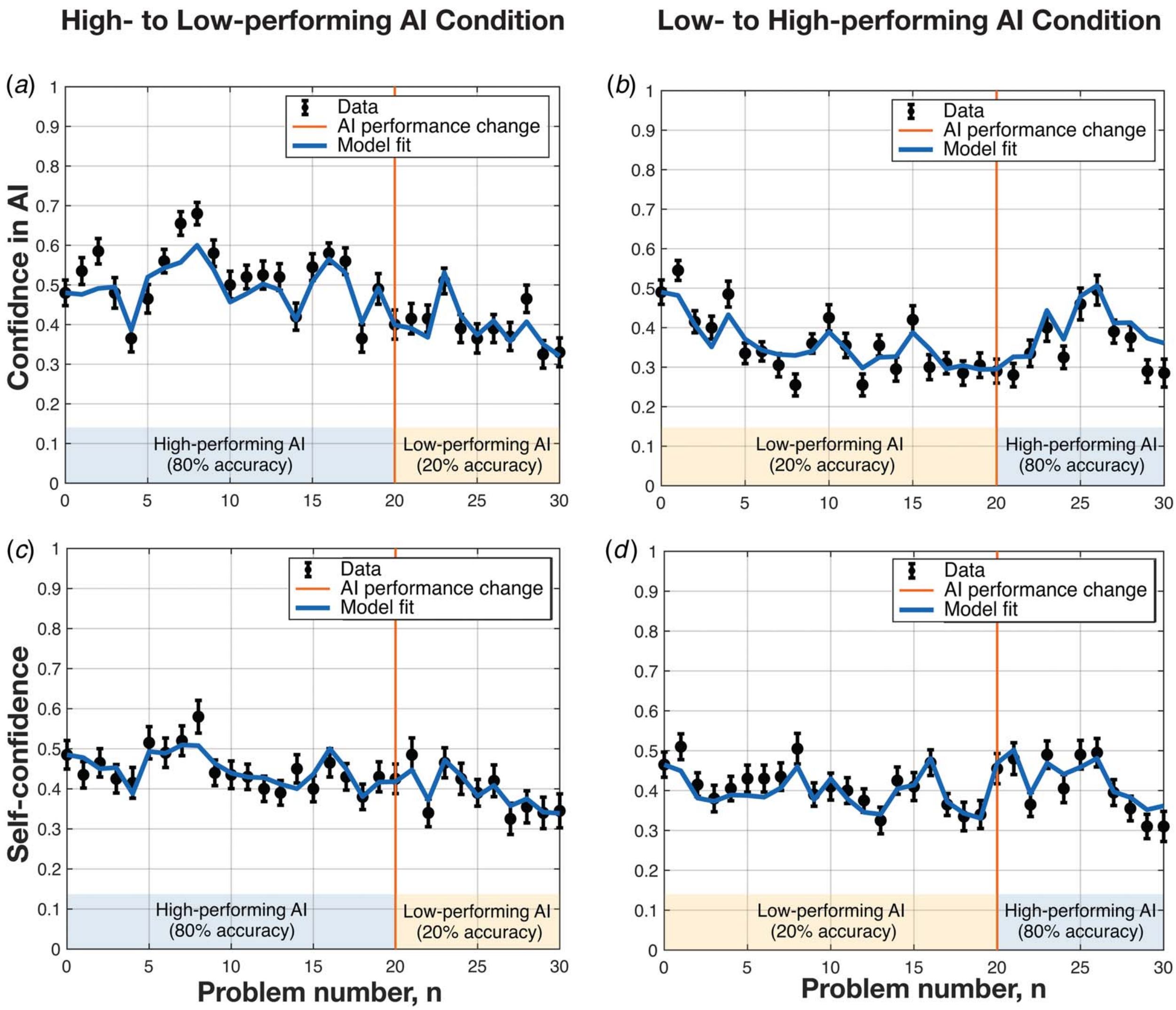


Fig. 3 Confidence in AI model fitting results in the (*a*) hi-lo condition and (*b*) lo-hi condition. Self-confidence model fitting results in (c) hi-lo condition and (*d*) lo-hi condition. Each pair (i.e., (*a*), (*b*) and (c), (*d*)) uses the same parameter values to fit the model to the data.

minor variations (below 0.1 mean absolute deviations from the values in Table 1). Therefore, the model parameter fitting results are robust.

With the optimal parameter values in Table 1 and the initial confidence values, the model iteratively calculates and predicts trial-by-trial human confidence in AI and human self-confidence in the experiment (see Fig. 3). The data points are the mean confidence of the 50 participants in the relevant condition. The standard error of the data is indicated by the error bars. The fitted lines show the mean model fitting results found using the optimal parameter values. Although the model computes confidence values at individual trials, line plots are used for a clear visual demonstration of the fit. The mean squared error (MSE) and adjusted R-squared value of the confidence in AI model (Figs. 3(*a*) and 3(*b*)) are 0.0017 and 0.75, respectively. The MSE and R-squared value of the self-confidence model (Figs. 3(*c*) and 3(*d*)) are 0.00076 and 0.75, respectively.

3 Results

3.1 Impact of Artificial Intelligence Accuracy and Its Change on Human Confidence. Figure 4 displays the confidence in AI and self-confidence results from the two experimental conditions. The data points are the mean confidence of the 50 participants in the relevant condition. The standard error of the data is indicated by the error bars. The linear fits before and after the AI performance change are also shown in the plots. The slopes of these fits are used to assess the rate and trend of change in confidence rather than the magnitude of change.

Figures 4(*a*) and 4(*b*) show the changes in designer's confidence in AI during the two conditions (hi-lo and lo-hi). When initially working with the AI before the AI performance change (i.e., problems 1 to 20), good AI performance in hi-lo condition does not significantly influence how confident designers are in the AI's ability (F-test, $p = 0.4$). However, poor AI performance in lo-hi condition decreases the participants' confidence in the AI (F-test, $p < 0.05$). After the change in the AI performance (i.e., after problem 20), the trend in the confidence in the AI switches in the same direction as the performance change, though not statistically significant (linear regression with interaction, $p = 0.5$ and 0.1, in hi-lo and lo-hi conditions respectively).

Figures 4(*c*) and 4(*d*) illustrate the self-confidence results. During the initial interactions with the AI (i.e., problems 1 to 20), both good (hi-lo condition) and bad (lo-hi condition) AI performances have a marginally significant, negative impact on self-confidence (F-test, $p = 0.06$ and 0.05, in hi-lo and lo-hi conditions respectively). With a switch in AI performance, regardless of the direction

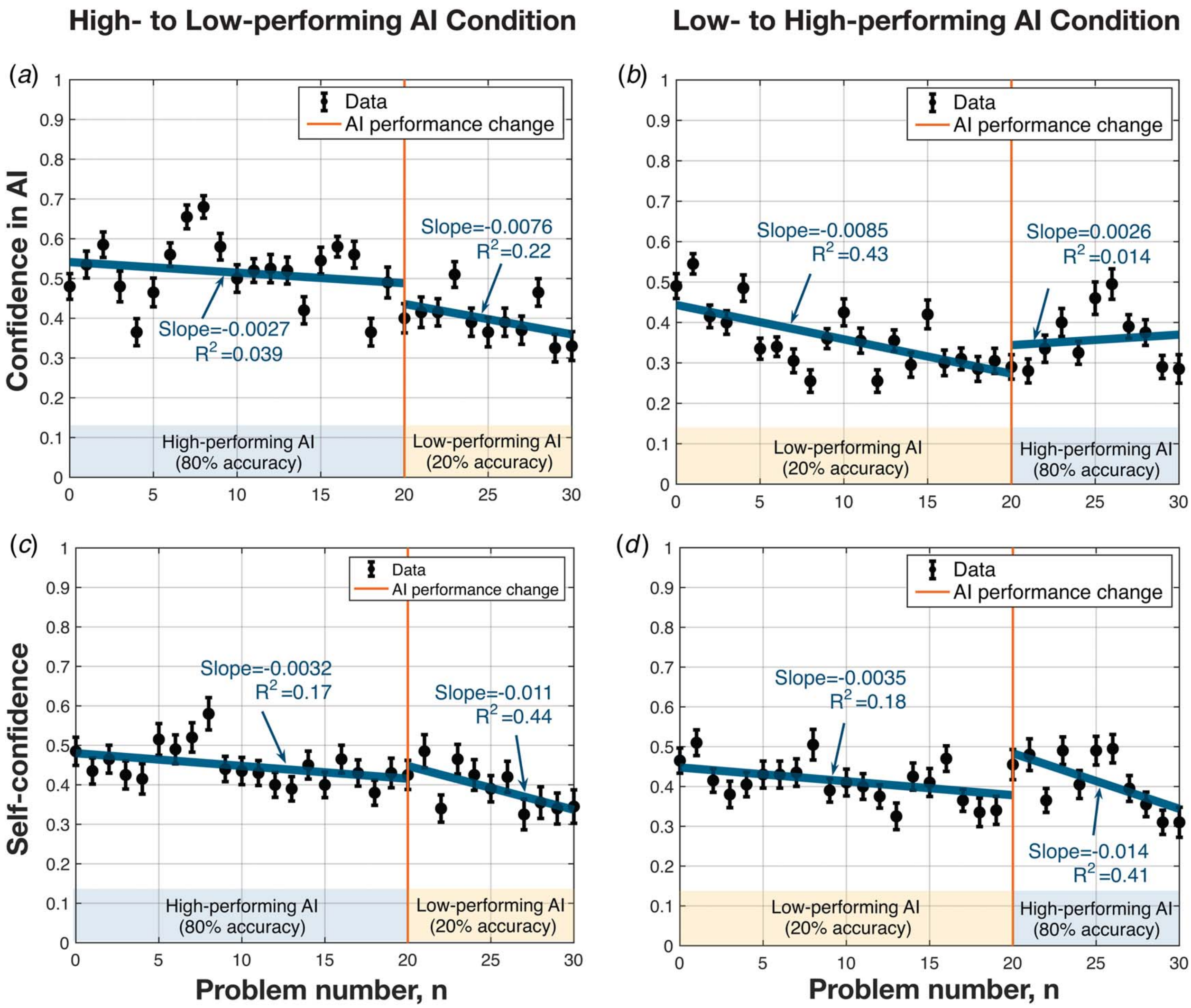


Fig. 4 Confidence in AI plots of the (*a*) hi-lo condition and (*b*) lo-hi condition. Self-confidence plots of the (*c*) hi-lo condition and (*d*) lo-hi condition.

of the switch, the trend of self-confidence changes in the negative direction. The magnitude of this change is marginally significant in both conditions (linear regression with interaction, $p = 0.1$ and 0.05, in hi-lo and lo-hi conditions respectively).

3.2 Impact of Different Types of Experiences on Human Confidence.

The estimated values of the parameters, ω_1, ω_2, ω_3, and ω_4, in the confidence model (see Table 1) illustrate the impact of the four types of experiences during AI-assisted decision-making on designers' confidence in the AI and in themselves. These results are repeated in Table 2 for convenience. As mentioned in the earlier description of the model, ω_1 and ω_3 correspond to the instances where designers are receiving either positive or negative feedback on the performance of the AI (i.e., accept AI suggestions) respectively. ω_2 and ω_4 correspond to those where designers are receiving either positive or negative feedback on their own performance (i.e., reject AI suggestions), respectively. For interpretation of the results in Table 2, ranging from zero to one, a value that is greater than 0.5 means the experience has positive impact on confidence, and a value that is less than 0.5 mean negative impact.

The results in the first row of Table 2 display the influence of the four experiences on the participant's confidence in the AI. Expectedly, positive and negative feedbacks on the performance of the AI respectively increase and decrease the participant's confidence in the AI ($\omega_1 = 0.870$ and $\omega_3 = 0.117$, respectively). Interestingly, however, any feedback on their own move, positive or negative, decreases designers' confidence in the AI ($\omega_2 = 0.252$ and $\omega_4 = 0.292$).

The results in the second row of Table 2 show how the four experiences influence the participants' self-confidence. As expected, positive and negative feedbacks on their own performance increase and decrease the participant's self-confidence ($\omega_2 = 0.833$ and $\omega_4 = 0.188$) respectively. When the feedback is about the AI's performance, however, positive feedback slightly increases designers' confidence in their own ability ($\omega_1 = 0.623$), while a negative one significantly decreases it ($\omega_3 = 0.195$).

Table 2 The optimal parameter values of the confidence model correspond to the impact of four types of experiences in AI-assisted decision-making on the participant's confidence in the AI and their self-confidence. 0.5 is neutral in that >0.5 is positive and <0.5 is negative

	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	0.870	0.252	0.117	0.292
Self-confidence	0.623	0.833	0.195	0.188

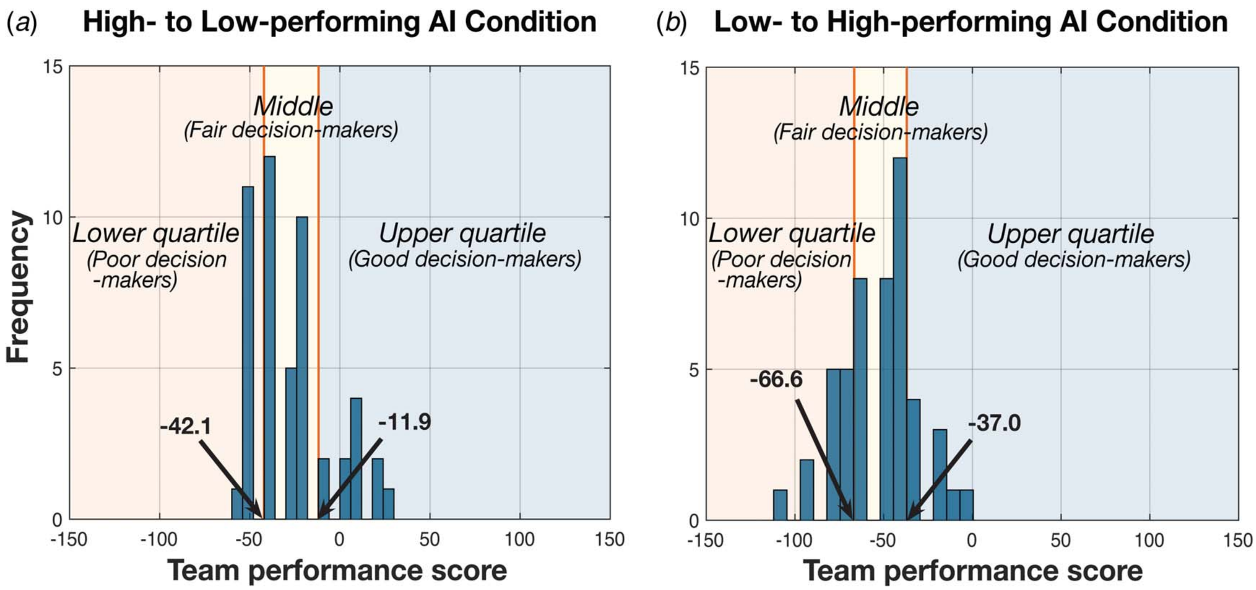


Fig. 5 Classification of poor, fair, and good decision-makers in the (*a*) hi-lo condition and (*b*) lo-hi condition, using the participants' final team performance scores

3.3 Impact of Human Confidence on Artificial Intelligence Acceptance Decisions.

Having found how designers' confidence in the AI and in themselves evolve over the course of teamwork due to dynamic AI performance, this section explores the impact of these confidences on their decision to accept or reject AI suggestions using logistic regression. There are two predictor variables which are designers' confidence in the AI and self-confidence. There is one binary response variable: decision to accept or reject AI suggestions; and there are 99 dummy variables that represent 100 subjects. The results reveal that confidence in the AI and self-confidence are both significantly correlated to the AI acceptance decisions (coefficient = −0.829 and 0.519, $p < 0.05$ and <0.05, respectively). Surprisingly, the more confident designers are in the AI's ability, the less likely they are to accept AI suggestions, and the more confident designers are in their own ability, the more likely they are to accept AI suggestions.

3.4 Characteristics of Successful Decision-Makers.

This section identifies decision-making patterns unique to successful decision-makers to gain insight into enhancing the outcome of AI-assisted decision-making. The participants are classified into three groups: poor, fair, and good decision-makers, based on their final team scores. For each participant, this final team score is calculated by summing the scores of the final decisions in the 30 problems, thereby representing how well the participant accepted or rejected the AI suggestions. First, for each of the two experimental conditions, 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 groups (see vertical lines in Fig. 5): lower quartile, middle two quartiles combined, and upper quartile, which correspond to poor, fair, and good decision-makers. Respectively, for the hi-lo condition (Fig. 5(*a*)), there are 12, 27, and 11 participants in these groups, while in the lo-hi condition (Fig. 5(*b*)), there are 13, 28, and 9 participants. Therefore, together, there are total 25, 55, and 20 participants categorized as poor, good, and fair decision-makers respectively. The results in Fig. 6 and Table 2 are constructed using such combined data. Each data point in the plots in Fig. 6 corresponds to one (or more participants if overlapped) participant.

Figure 6(*a*) shows the relationship between the participants' independent skill level and the final human-AI team score. Again, final team score represents goodness of human designer's decision-making ability. There is some overlap in the final scores among poor, fair, and good decision-makers because the plot combines results from the two experimental conditions that each has its own upper and lower quartile thresholds. The results in Fig. 6(*a*) demonstrate that in this experiment, designers' decision-making skills do not reflect their truss design skills. The three levels of decision-makers show a very similar range of individual performance scores between −100 and 0. Additionally, Fig. 6(*b*) displays the relationship between the participant's decision-making ability (final team performance) and their confidence in their own design ability. The results show that self-confidence of the three levels of decision-makers varies over a similar range, in the same manner as their individual skills. Therefore, successful decision-makers do not show a distinct level of design ability or confidence in their own design ability.

Then, Fig. 7 illustrates the differences in the probability of accepting AI suggestions among the varying levels of decision-makers. It is observed that good decision-makers in the hi-lo condition have a distinctly high probability of accepting AI suggestions while those in the lo-hi condition have a distinctly low probability. This difference in the probability between the two conditions among the good decision-makers expectedly demonstrates that they can appropriately determine when to accept or reject AI suggestions. Although this difference also exists among the poor and fair decision-makers, it is much smaller, meaning that they are not making as drastic changes to their acceptance rate according to the observed AI accuracy as the good decision-makers.

Now, a regression analysis is conducted to examine how confidence in AI and/or self-confidence of the good decision-makers affects their decisions to accept or reject AI suggestions differently from other decision-makers. First, as shown in the first row of Table 3, all three groups demonstrate negative correlation between designers' confidence in the AI and their probability of accepting AI suggestions (coefficient = −0.729, −0.848, and −1.05, $p = 0.05$, <0.05, and = 0.05, in the order of poor to good decision-makers). Generally, self-confidence is positively correlated with designers' likelihood of accepting AI suggestions (coefficient = 0.895, 0.191, and 1.07 for poor to good decision-makers), however with varying levels of significance among the three groups. This relationship is significant among poor decision-makers, not significant among fair decision-makers, and marginally significant among good decision-makers ($p < 0.05$, = 0.5, and = 0.07, respectively).

3.5 Post-Experiment Questionnaire.

The responses to the five questions in the post-experiment questionnaire (see Fig. 8)

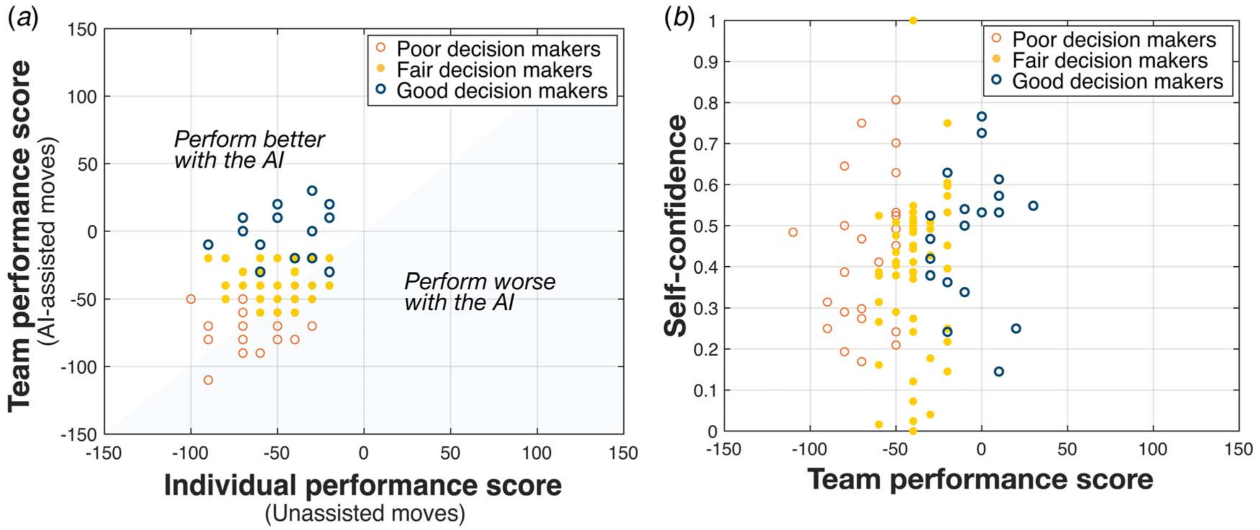


Fig. 6 Participants' team performance score (decision-making ability) versus (*a*) individual performance (task ability) and (*b*) confidence in their own task ability

are used to further understand the results of the experiment as elaborated in Sec. 4. The first two questions are designed to learn about the participants' perception of the AI, while the next two attempted to learn about their perception of themselves. Then, the final question asks about their perception of their own decision-making process. The responses from the two conditions are compared. Again, in the hi-lo condition, AI performance changes from 80% to 20% accuracy two-thirds of the way through the experiment, while in the lo-hi condition, it changes from 20% to 80% accuracy.

3.5.1 How Helpful Were the Artificial Intelligence Suggestions in Doing This Task?. 38% of the participants from the hi-lo condition think the AI suggestions were helpful or very helpful, and 32% of them think they were unhelpful or very unhelpful. In contrast, in the lo-hi condition, only 6% of the participants think the AI suggestions were helpful or very helpful, and 58% of them think they were unhelpful or very helpful. These results may be suggesting that the first impression (initial AI performance) counts.

3.5.2 Was the Quality of the Artificial Intelligence Suggestions Consistent? If Not, How Did it Change?. It is important to note that this question is dependent on how often designers accepted AI suggestions to gain information about AI performance. A big percentage of participants from both the hi-lo and lo-hi conditions (44% and 48% respectively) answered "I am not sure" and additional 6% and 24% answered "Yes, it was consistent" to this question. The discrepancy between the two conditions in the number of participants who correctly identified the change in AI accuracy is very interesting. While 48% of the participants in the hi-lo condition correctly reported that the AI got worse over time, only 16% did in the lo-hi condition. These responses together show that in the hi-lo condition, there is a dichotomy between those who could and could not identify the change in AI performance, while in the lo-hi condition, very few could accurately identify the change.

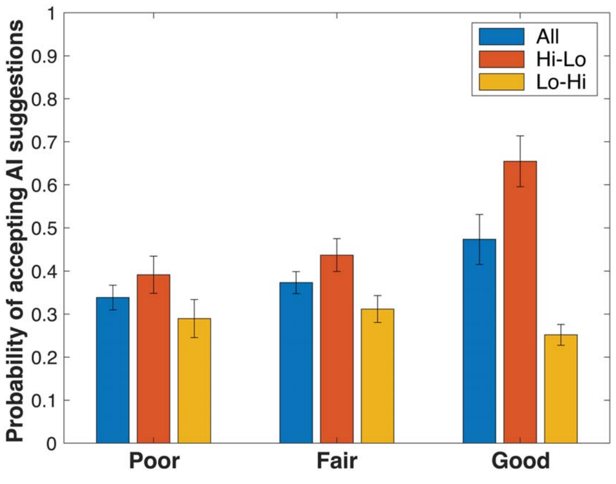


Fig. 7 Probability of accepting AI suggestions among poor, fair, and good decision-makers

3.5.3 How Good Were You at Designing Trusses?. The hi-lo and lo-hi conditions show similar responses for this question. In both conditions, the largest percentage of participants answer that they were "bad" at designing trusses, which is then followed by "very bad". There are less than 10% of participants think they were good or very good at designing trusses.

3.5.4 How Good Were You at Making the Final Decision of Which Move to Choose Between Your Own Move and an Artificial Intelligence Suggestion?. While the biggest percentages of participants in both conditions think that they were bad at making the final decisions, just like in question 3, notably fewer participants perceive themselves as "very bad". Generally, compared to judging how well they were at designing trusses, the participants are more positive about how well they made the final decisions.

3.5.5 When Deciding Between Your Own Design Move and an AI Suggestion, What Did You Consider More: AI's Ability to Design Trusses or Your Own Ability to Design Trusses?. In the hi-lo condition, 32% of the participants report that they considered the AI's ability more, and another 32% reported that they considered the AI's ability and their own ability equally. In contrast, in the lo-hi condition, 60% of the participants report that they considered their own ability more, while only 14% and 16% respectively report "AI's ability" and "both equally".

4 Discussion

The discussion elaborates on the major results of this work and compares them to the results from Chong et al. [21]. The purpose of Chong et al. [21] was to understand how human confidence in AI and self-confidence evolve and impact AI acceptance decisions during AI-assisted decision-making. To achieve this goal, Chong

Table 3 Regression results between designers' confidence in AI, self-confidence, and their probability of accepting AI suggestions for varying levels of decision-makers

		Decision-making skill level		
		Poor	Fair	Good
Regression against the probability of accepting AI suggestion	Confidence in AI	−0.729 ($p = 0.05$)	−0.848 ($p < 0.05$)	−1.05 ($p = 0.05$)
	Self-confidence	0.895 ($p < 0.05$)	0.191 ($p = 0.5$)	1.07 ($p = 0.07$)

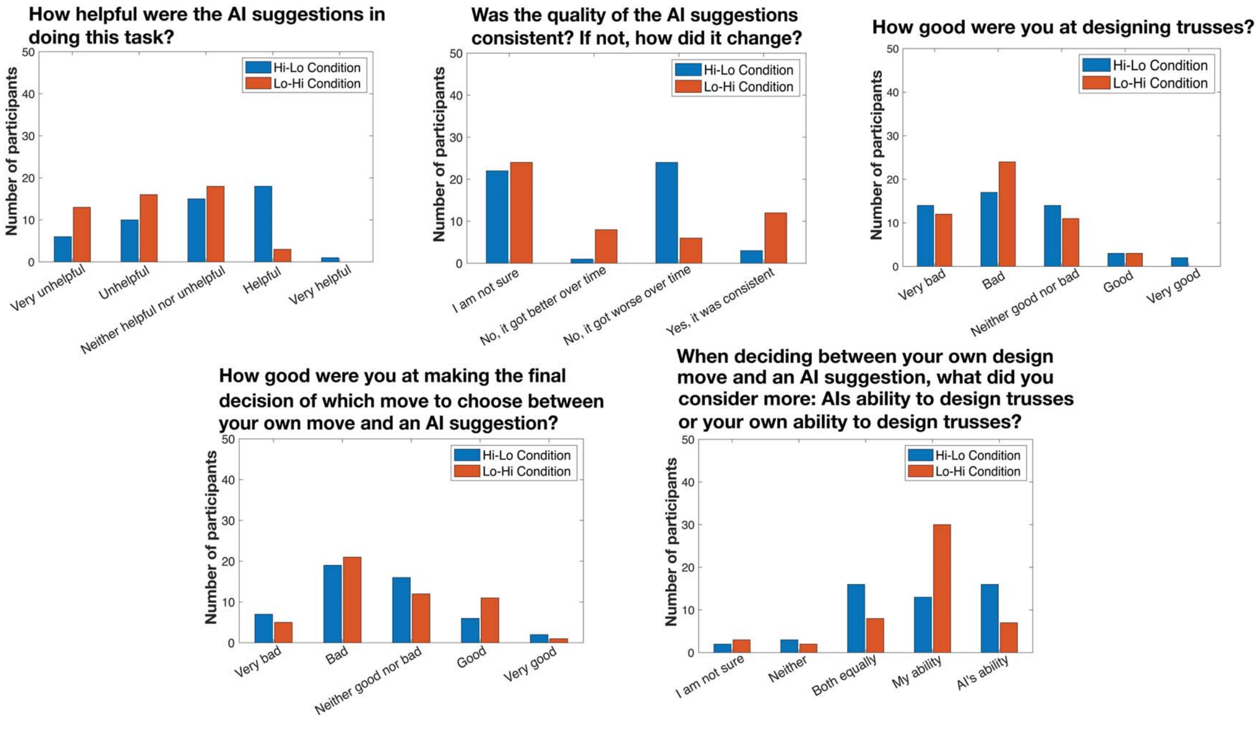


Fig. 8 Responses from the post-experiment questionnaire

et al. [21] used a chess puzzle task, in which given a chess board state, participants needed to make the best next move. As with the current study, they first selected a move independently without an AI suggestion, then made a final move decision after receiving an AI suggestion. For each final decision, feedback was provided depending on the quality of the move. Although Chong et al.'s [21] work provided much insight, the design of the study limits the results to be directly applicable only to AI-assisted decision-making scenarios with a chess puzzle task. It is unknown which of and how their results generalize to or are relevant in design settings. Therefore, the current work aims to study the evolution and influence of human confidence in AI and self-confidence specifically during AI-assisted decision-making in design. For this goal, the experiment in this work is designed to resemble Chong et al. [21], except for the task, which is the truss design task.

4.1 Differences Between the Chess Puzzle Task and the Truss Design Task. Despite some similarities such as task procedure (i.e., unassisted then assisted action) and discreteness of problems (i.e., each problem does not depend on others), the truss design task and the chess puzzle task have many differences that may produce inconsistent results between the current study and Chong et al. [21]. One of these differences is that AI systems are prevalently known to be highly proficient at solving chess problems while perhaps not quite as known for their proficiency in designing. In chess, there are universally accepted, well-known AI algorithms such as Stockfish that perform better than human experts, but such algorithms do not exist yet in design, specifically in design of trusses. Such assumptions about AI systems' skills in these settings could influence participants' confidence and decisions throughout the teamwork. Secondly, the truss design task does not have an opponent while chess puzzle task does, possibly bringing about different considerations during decision-making. Although all problems are independent in both tasks, when working on chess problems, the presence of an opponent is implied due to the nature of the game of chess. Therefore, although a design action is always reversible under the control of the designer in the truss design task, a chess move is less so because the opponent may make an action that prevents a reversal. This reversibility of the actions in the truss design task may place less importance on each action the participants take than in the chess puzzle task where actions are not always reversible. Third, the impact of truss design actions is less differentiable than that of chess moves. Because the objective of the truss design task to maximize the SWR consists of two conflicting aspects: one to make the truss stronger and another to make it light, a design action may be beneficial in one aspect but not in another. This characteristic of the truss design task may lead to difficulties in determining the overall impact of AI suggested actions and therefore in deciding whether to accept or reject them. This difficulty is hinted at by the team and individual performance scores in Fig. 6 which are both concentrated in lower ranges than the scores in Chong et al. [21]. Considering the discrepancies listed above, the comparison of the results in this study to those from Chong et al. [21] can reveal insights into which findings are potentially generalizable across different tasks and which are unique to each task. Most importantly for this paper, the comparison shows which results about confidence in AI and self-confidence applies exclusively to design situations. For each research question, the pertinent results from this study are first discussed, followed by their comparison to the results from Chong et al. [21].

4.2 Major Results From the Current Work

4.2.1 Decreasing Trends in Designers' Confidence. The current work first answers how AI performance and its variations affect designers' confidence in AI and self-confidence during AI-assisted decision-making in design, given prior proposals that human confidence is prone to change based on the performance of the AI [25–28]. The results in this work reveal that designers initially have relatively high confidence in the AI, which quickly decreases with poor AI performance but does not increase much with good AI performance. Self-confidence, however, is hardly affected by the initial level of AI performance. Then, when the AI accuracy changes, regardless of the direction of change, designers become slightly less confident in themselves but remain unaffected in their confidence in the AI. This directional independence may be explained by the result from the post-experiment questionnaire that many participants did not accurately identify the change in the AI performance. Additionally, in this experiment, neither confidence in the AI nor self-confidence increases even with good AI performance. This lack of impression of good AI performance is also evident in the responses from the post-questionnaire (Sec. 3.5). In question 1, 32% of the participants from the hi-lo condition reported the AI to be unhelpful, even though they mainly interacted with a proficient AI throughout the experiment. In question 2, 84% of the participants in the lo-hi condition could not identify the improvement in the AI performance. Overall, these insights provide information about how different AI accuracies influence human confidence in AI and self-confidence and can be helpful in deploying AIs with appropriate proficiencies into design teams.

4.2.2 Mis-Inference of Information From Feedback. The results also demonstrate that designers tend to infer faulty information from feedback in this study. First, when designers reject the AI suggestion and receive feedback on their own performance, this feedback affects their confidence in AI, although it does not imply any information about the AI's ability. Both positive and negative feedback on their own performance result in lower confidence in the AI. This may be a manifestation of human self-serving bias, which is the individual's tendency to "attribute success to their own personal dispositions and failure to external forces" [29,30]. When the participants from the study are told that their action is good, they may attribute this success solely to themselves and in turn lose confidence in the AI. In contrast, when the participants are told that their action is bad, they may attribute some of this failure to the AI. Erroneous inference of information also occurs when the participants receive feedback on the AI's performance. Despite the absence of information about their own ability, the feedback on the AI's performance alters designers' self-confidence in the same direction as the feedback. This may be favorable from a managerial perspective if designers are taking responsibility as the final decision-maker. However, such attribution of responsibility could hinder human-AI team performance when designers use inaccurate information to make decisions. The aforementioned findings about the impact of different feedback on human confidence enable detection and reasoning of changes in the levels of confidence in the AI and in themselves during the design process.

4.2.3 Strong Correlation Between Designers' Confidences and Their Acceptance of Artificial Intelligence Suggestions. The second research question then asks how designers' confidence in the AI and self-confidence are associated with the probability of accepting AI's design suggestions. The results show that both confidences are strongly correlated to the AI acceptance decisions. First, when people are more confident in themselves, they are more likely to accept AI suggestions. This result can be understood by the earlier results that the participants' confidence in their own ability increases only when they receive positive feedback on either themselves or the AI. This feedback provides neutral to positive information about the AI, therefore increasing the participants' likelihood of accepting AI input. Similarly, with lower self-confidence, the participants are less likely to accept AI suggestions. Second, it is found that the more confident designers are in the AI's ability, the less likely they are to accept AI suggestions. Borrowing earlier results, the confidence in the AI increases only when the participants receive positive feedback on the AI's performance, meaning that oddly, the positive feedback on the AI decreases the chance of accepting AI input. Though indefinite, the participants may be making such suboptimal decisions when accepting or rejecting AI suggestions because the impact of the AI suggestions is not clearly differentiable in the truss design task. These insights into how designers' confidence in AI and self-confidence are related to the decision to accept or reject AI suggestions shine light on their cognitive processes during decision-making and inspire ways to enhance the effectiveness of AI-assisted decision-making in design.

4.2.4 Characteristics of Successful Decision-Makers. Finally, the characteristics of poor, fair, and good decision-makers are compared to identify the unique patterns in good decision-making. There are no differences in the participants' truss design skills (summed score of the unassisted actions) or their confidence in their own design skills (average reported self-confidence) among the three groups, meaning that these characteristics are not what leads to good decision-making. However, good decision-makers show a greater difference in their probability of accepting AI suggestions between the two conditions (hi-lo and lo-hi) than poor and fair decision-makers, meaning that they are more appropriately adjusting their acceptance rate according to the observed AI accuracy. Despite this unique characteristic of good decision-makers, their confidence impact the probability of accepting AI suggestions in a similar manner among the three levels of decision-makers. All three groups display a negative correlation between the participant's confidence in the AI and their probability of accepting AI suggestions, just as in the combined regression result. Self-confidence has a negative relationship with the likelihood of AI acceptance among poor and good decision-makers, while fair decision-makers distinctively show no significant relationship.

4.3 Comparison of the Results to Chong et al.'s Chess Study (2022). In comparison to the results in Chong et al. [21] that used the chess puzzle task instead of the truss design task, initial AI performance, both good and bad, affects human confidence in the same manner. In the initial stages of human-AI teamwork, there are many results that are consistent across both tasks such as the high starting confidence in AI, the constant level of self-confidence, and the negative impact of low AI accuracy on human confidence in AI. However, the change in AI performance has very different effects on the participants' confidence in AI and their self-confidence depending on the task. In Chong et al. [21], human confidence in AI shifted in the same direction as the AI performance change, and self-confidence was only affected by a negative change in the AI performance. In contrast, in the current study, confidence in the AI is unaffected, and self-confidence marginally decreases independent of the direction of the AI performance change. The lack of influence of the switch in AI accuracy on the participant's confidence in the AI in design, unlike in chess, may be because this switch is not as easily identifiable in the truss design task. This reason is supported by the result from the post-experiment questionnaire that many designers did not correctly perceive the accuracy of the AI and its change. Furthermore, while the participants from Chong et al. [21] showed decrease in self-confidence only with poor AI performance, the participants from the current work lost self-confidence even with good AI performance. During the truss design task, the participants may plainly be getting discouraged by the indistinct impact of the AI suggested actions.

Most of the results about the inference of information from feedback look the same as in Chong et al. [21], except when humans receive negative feedback on their own performance. This means that independent of the tasks, the experiences (and their corresponding feedback) during AI-assisted decision-making mostly affect human confidence in the AI and self-confidence in a similar manner. However, negative feedback on human performance decreased the participants' confidence in the AI during the truss design task but had little to no effect during the chess puzzle task. This difference could be because the cost and benefit of AI suggested actions in the truss design task are less differentiable than in the chess puzzle task. Learning about their own poor performance in the task while perceiving that the goodness of design actions is not obvious, the participants may expect the AI to find the task difficult, lowering their confidence in the AI.

Generally, the participants in the current study show lower or more easily decreasing self-confidence and confidence in AI than those in the chess study. It is important to note that evidently from the participants' data, the less differential nature of the impact of truss design actions may be making the task more difficult than the chess puzzle task. Most participants in this study received a final score below ten (i.e., 16 disadvantageous and 14 advantageous actions), which means that they made more disadvantageous moves than advantageous ones during the experiment. The participants in the chess study, however, mostly received a final score below 40 (i.e., 11 disadvantageous and 19 advantageous moves), which means more of them made more advantageous moves than disadvantageous ones than those in the truss study. This lower success rate in the truss task may explain the lower and decreasing trend of designers' self-confidence and confidence in AI in this study compared to the participants' confidence in the chess study. This potential relationship between the success rate and self-confidence is also supported by the post-experiment questionnaire that there are less than 10% of participants who thought they were good or very good at designing trusses. Despite the possible relationship between the success rate on the confidence levels, the influence of AI performance on designers' confidence remains supported as the results in Table 1 show that even with the difficulty of the truss design task, e_1 increases designers' confidence in AI and self-confidence, and e_2 increases their self-confidence.

The surprising results about how human confidence in AI and/or self-confidence are correlated to the decisions to accept or reject AI input run counter to those from Chong et al. [21]. Although Chong et al. [21] showed that only self-confidence, not confidence in the AI, is significantly related to AI acceptance decisions in chess, both confidences are related to the decisions in the current work on the design. This difference may be explained by the characteristic of the truss design task that the impact of the suggested design actions is less clear than in the chess puzzle task. When the goodness of the AI suggestions is not as obvious, designers may consider their confidence in the AI's ability from earlier interactions more during their decision-making. Unfortunately, considering their confidence from prior interactions seems to be leading to undesirable decisions: the higher the confidence in the AI, the less likely they are to accept the AI input. This relationship between designers' confidence in AI and their acceptance of AI suggestions is in the opposite direction from the chess study. Additionally, human self-confidence is associated with the AI acceptance decisions also in the opposite direction from that of Chong et al. [21]. In this work, the more the participants are confident in their own ability, the more likely they are to accept the AI's suggestion, while in chess, the participants were less likely to. These different results may again be because of the indistinct cost and benefit of AI suggestions in the truss design task, which could be impeding the appropriate transfer of their confidence levels to AI acceptance decisions and resulting in unexpected behaviors.

Finally, although the results from this work do not show many unique characteristics among successful decision-makers, Chong et al. [21] found a vicious cycle that they tend to avoid, achieving better results than others. In chess, poor decision-makers fall into this vicious cycle when interacting with an unskilled AI where they repeatedly rely on this AI because they attribute blame to themselves (decrease in self-confidence), consequently accepting the next AI suggestion again. However, good decision-makers do not enter this cycle because, with decreases in self-confidence, they are less likely to accept the next AI suggestion. In the present work, like the good decision-makers in Chong et al. [21], all levels of decision-makers show a positive correlation between self-confidence and the probability of accepting AI suggestions, therefore avoiding the vicious cycle.

4.4 Implications for Design Practice. The insights from this work can significantly impact the practice of AI-assisted decision-making in design. First, this work involves much information about how human confidences change with individual experiences in AI-assisted decision-making. Such information allows maneuvering of experiences to calibrate human confidence in AI and in them. For example, when designers are suffering from low self-confidence that may lead to inappropriate reliance on AI input, experiences that would increase self-confidence might be provided repeatedly to increase their self-confidence. Second, the detrimental impact of poorly performing AIs on human confidence shown in this work may increase the accuracy threshold for including AIs in design teams. Additionally, the results together conclude that AI is not a panacea for design problem-solving. Even with high AI performance, human-AI teams may not reach the desired design performance level without successful management of human designers' confidence levels. According to the results from this work, some challenges unique to design may include preventing the drop in designers' self-confidence and effectively communicating the quality of AI suggestions (especially when it is good) to designers.

4.5 Limitations and Areas of Future Work. There are some limitations of this work that provide opportunities for future work. First, the fit of the dynamic model of human confidence can be improved. The R-squared value of 0.75 is sufficient for the purpose of this study which is to understand the impact of different types of experiences on the change in human confidence levels. To be used for confidence calibration applications, it is beneficial to increase the model accuracy by including other factors of human confidence in the model or by testing different forms of relationship between the different factors. A second limitation is the broad consideration of AI in this work. This work distinguishes AI from other types of computational agents by its data-driven approach. For more AI-specific insights, it will be helpful to explore distinctive properties of AI systems that best align with designers' confidence during problem-solving.

It is important to recognize that the findings may not be generalizable to all design scenarios because this work only covers one specific type of design problem. The results can be broadly applied to short-term, well-defined design decision-making contexts where a human designer regularly receives AI input. Long-term AI-assisted decision-making scenarios, where there are breaks with no interaction between humans and AIs, may show differences in human confidence and decision-making and therefore may not be a context that this work's results are directly applicable. As long as the quality of each design decision can be measured and an AI regularly provides design-related inputs, the insights from this work should help improve the outcome of AI-assisted decision-making.

5 Conclusion

This work conducts a cognitive study and leverages a quantitative model to investigate the changes in designers' confidence in AI and their self-confidence during AI-assisted decision-making in design and how these confidences affect their decisions. The results demonstrate that during AI-assisted decision-making in design, designers initially are highly confident in their AI teammates but quickly become less confident with poor AI performance. Their confidence in their own ability, however, does not change with initial AI performance. This paper also shows that when the AI accuracy changes, the trend of designers' self-confidence alters negatively independent of the direction of the change while the trend of their confidence in the AI remains the same as it was before the change. Furthermore, throughout the teamwork, designers are inclined to deduce incorrect information from feedback, judging the AI's or their own ability based on the feedback given to the other. Considering such dynamic changes in confidence of designers, confidence in the AI and self-confidence are positively and negatively (respectively) correlated to their chance of accepting an AI suggestion. Moreover, this work presents many parallels, rather than discrepancies, in the decision-making characteristics between different levels of decision-makers. Finally, this work identifies the findings that are unique to design, as well as those that may be generalizable across different types of tasks.

Acknowledgment

This work was supported by the Air Force Office of Scientific Research (AFOSR) under grants FA9550-18-1-0088 and FA9550-21-1-0442. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsor.

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

References

[1] Chen, H. Q., Honda, T., and Yang, M. C., 2013, "Approaches for Identifying Consumer Preferences for the Design of Technology Products: A Case Study of Residential Solar Panels," ASME J. Mech. Des., **135**(6), p. 061007.

[2] Camburn, B., Arlitt, R., Anderson, D., Sanaei, R., Raviselam, S., Jensen, D., and Wood, K. L., 2020, "Computer-Aided Mind Map Generation Via Crowdsourcing and Machine Learning," Res. Eng. Des., **31**(4), pp. 383–409.

[3] Williams, G., Meisel, N. A., Simpson, T. W., and McComb, C., 2019, "Design Repository Effectiveness for 3D Convolutional Neural Networks: Application to Additive Manufacturing," ASME J. Mech. Des., **141**(11), p. 111701.

[4] Nie, Z., Lin, T., Jiang, H., and Kara, L. B., 2021, "TopologyGAN: Topology Optimization Using Generative Adversarial Networks Based on Physical Fields Over the Initial Domain," ASME J. Mech. Des., **143**(3), p. 031715. .

[5] Zhang, W., Yang, Z., Jiang, H., Nigam, S., Yamakawa, S., Furuhata, T., Shimada, K., and Kara, L. B., 2019, "3D Shape Synthesis for Conceptual Design and Optimization Using Variational Autoencoders," Proceedings of the IDETC/CIE, Anaheim, CA, Aug. 18–21, ASME Paper No. DETC2019-98525.

[6] Raina, A., McComb, C., and Cagan, J., 2019, "Learning to Design From Humans: Imitating Human Designers Through Deep Learning," ASME J. Mech. Des., **141**(11), p. 111102.

[7] Raina, A., Puentes, L., Cagan, J., and McComb, C., 2021, "Goal-Directed Design Agents: Integrating Visual Imitation With One-Step Lookahead Optimization for Generative Design," ASME J. Mech. Des., **143**(12), p. 124501.

[8] Lopez, C. E., Miller, S. R., and Tucker, C. S., 2019, "Exploring Biases Between Human and Machine Generated Designs," ASME J. Mech. Des., **141**(2), p. 021104.

[9] Song, B., Zurita, N. F. S., Zhang, G., Stump, G., Balon, C., Miller, S. W., Yukish, M., Cagan, J., and McComb, C., 2020, "Toward Hybrid Teams: A Platform to Understand Human-Computer Collaboration During the Design of Complex Engineered Systems," Proceedings of the Design Society: DESIGN Conference, Virtual, Oct. 26–29, pp. 1551–1560, 1.

[10] Wilson, H. J., and Daugherty, P. R., 2018, "Collaborative Intelligence: Humans and AI are Joining Forces," Harv. Bus. Rev., 96(4), pp. 114–123

[11] Zhang, G., Raina, A., Cagan, J., and McComb, C., 2021, "A Cautionary Tale About the Impact of AI on Human Design Teams," Des. Studies, **72**, p. 100990.

[12] Lee, J. D., and See, K. A., 2004, "Trust in Automation: Designing for Appropriate Reliance," Human Factors, **46**(1), pp. 50–80.

[13] Parasuraman, R., and Riley, V., 1997, "Humans and Automation: Use, Misuse, Disuse, Abuse," Human Factors, **39**(2), pp. 230–253.

[14] Zhang, Y., Vera Liao, Q., and Bellamy, R. K. E., 2020, "Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making," Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, Jan. 27–30, pp. 295–305.

[15] Richtel, M., and Dougherty, C., 2015, "Google's Driverless Cars Run into Problem: Cars With Drivers," New York Times, September 2, 2015. https://www.nytimes.com/2015/09/02/technology/personaltech/google-says-its-not-the-driverless-cars-fault-its-other-drivers.html.

[16] Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E., 2019, "Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance," Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Skamania, WA, Oct. 28–30, pp. 2–11, 7.

[17] Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., and Horvitz, E., 2019, "Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff," Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, Jan. 27–Feb. 1.

[18] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P., 2003, "The Role of Trust in Automation Reliance," Int. J. Human Comput. Stud., **58**(6), pp. 697–718.

[19] Hoffman, R. R., Johnson, M., Bradshaw, J. M., and Underbrink, A., 2013, "Trust in Automation," IEEE Intell. Syst., **28**(1), pp. 84–88.

[20] Siau, K., and Wang, W., 2018, "Building Trust in Artificial Intelligence, Machine Learning, and Robotics," Cutter Business Technol. J., **31**(2), pp. 47–53.

[21] Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., and Cagan, J., 2022, "Human Confidence in Artificial Intelligence and in Themselves: The Evolution and Impact of Confidence on Adoption of AI Advice," Comput. Human Behav., **127**, p. 107018.

[22] Mayer, R. C., Davis, J. H., and Schoorman, F. D., 1995, "An Integrative Model of Organizational Trust," Acad. Manage. Rev., **20**(3), pp. 709–734.

[23] Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C., 1998, "Not So Different After All: A Cross-Discipline View of Trust," Acad. Manage. Rev., **23**(3), pp. 393–404.

[24] McComb, C., Cagan, J., and Kotovsky, K., 2015, "Rolling With the Punches: An Examination of Team Performance in a Design Task Subject to Drastic Changes," Des. Studies, **36**, pp. 99–121.

[25] Hu, W. L., Akash, K., Reid, T., and Jain, N., 2019, "Computational Modeling of the Dynamics of Human Trust During Human-Machine Interactions," IEEE Trans. Human-Machine Syst., **49**(6), pp. 485–497.

[26] Moré, J. J., and Sorensen, D. C., 1983, "Computing a Trust Region Step," SIAM J. Sci. Statist. Comput., **4**(3), pp. 553–572.

[27] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., and Parasuraman, R., 2011, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," Human Factors, **53**(5), pp. 517–527.

[28] Schoorman, F. D., Mayer, R. C., and Davis, J. H., 2007, "An Integrative Model of Organizational Trust: Past, Present, and Future," Acad. Manage. Rev., **32**(2), pp. 344–354.

[29] Campbell, W. K., and Sedikides, C., 1999, "Self-Threat Magnifies the Self-Serving Bias: A Meta-Analytic Integration," Rev. Gen. Psychol., **3**(1), pp. 23–43.

[30] Larson, J. R., 1977, "Evidence for a Self-Serving Bias in the Attribution of Causality," J. Pers., **45**(3), pp. 430–441.