A cautionary tale about the impact of AI on human design teams

Guanglu Zhang, Ayush Raina, Jonathan Cagan and *Christopher McComb*,
Department of Mechanical Engineering, Carnegie Mellon University, USA,
School of Engineering Design, Technology, and Professional Programs, The Pennsylvania State University, USA

Recent advances in artificial intelligence (AI) offer opportunities for integrating AI into human design teams. Although various AIs have been developed to aid engineering design, the impact of AI usage on human design teams has received scant research attention. This research assesses the impact of a deep learning AI on distributed human design teams through a human subject study that includes an abrupt problem change. The results demonstrate that, for this study, the AI boosts the initial performance of low-performing teams before the problem change but always hurts the performance of high-performing teams. The potential reasons behind these results are discussed and several suggestions and warnings for utilizing AI in engineering design are provided.

Keywords: artificial intelligence, collaborative design, engineering design, human-computer interaction, problem solving

Recent advances in artificial intelligence (AI), especially machine learning, enable researchers to develop various AI agents (referred to here simply as AIs) to support engineering design. AI has been applied in multiple phases of the engineering design process, including but not limited to, customer preference identification (Chen, Honda, & Yang, 2013), concept generation (Camburn, Arlitt, et al., 2020; Singh & Gu, 2012), concept evaluation (Camburn, He, Raviselvam, Luo, & Wood, 2020), prototyping (Dering, Tucker, & Kumara, 2017), and manufacturing (Williams, Meisel, Simpson, & McComb, 2019). Research results show that a well-trained AI is able to perform a specified design task as good as, or sometimes even better than, human designers (Lopez, Miller, & Tucker, 2019; Raina, Cagan, & McComb, 2019). For example, 2D boat sketches generated by a trained deep generative model are more likely to float and move in a computer simulation environment than human sketches on average (Lopez et al., 2019).

Corresponding author:
Guanglu Zhang
glzhang@cmu.edu

www.elsevier.com/locate/destud
0142-694X *Design Studies* **72** (2021) 100990
https://doi.org/10.1016/j.destud.2021.100990

Although these research results involving AI are promising, human designers treat AI as a design tool rather than their teammate, at least for now. As a powerful tool that supports human design, AI is not able to replace the human at the current stage because human creativity and agility are in need for solving complex and volatile engineering design problems (Song et al., 2020). The combination of these human characteristics with the high task-specific performance of AIs may be particularly effective. A recent study involving 1500 companies shows that the performance of a company improves significantly only when human and AI work together (Wilson & Daugherty, 2018). In the light of the growing applications of AI in engineering design, it is critical to understand how AI affects humans in the design process.

Although teaming is a prevalent approach to engineering design, human-AI interaction is under-studied in this context. Prior research endeavors in the engineering design community involving AI focus on developing AI for various design tasks rather than investigating the impact of AI on human design teams.

This research assesses the impact of a deep learning AI on the performance, behavior, and perceived workload of distributed design teams through a human subject study. An abrupt design problem change is included in the study to investigate whether the AI helps human design teams adapt to such change. The human confidence in AI and the human self-confidence in solving the design problem are also investigated. This paper provides new insights for distributed teams working with a deep learning AI to solve an engineering design problem.

In the human subject study, teams of three solve the same configuration design problem through a graphical user interface (GUI) on their own or with the availability of the deep learning AI to advise them. Halfway through the study, the design problem abruptly changes, requiring the team to adapt to the new problem goals. The significance of the use of distributed teams both highlights the use of technology for communication and design and enables experimental control and extensive data collection. Of note, verbal communication is not allowed in the study, but participants communicate directly by sharing designs with their teammates through the GUI.

The deep learning AI (Raina, McComb, & Cagan, 2019) is trained over data collected from a previous human design study that uses the same GUI (McComb, Cagan, & Kotovsky, 2015) but does not include learning over the changed problem. Since it is impossible to train a deep learning AI on every possible design constraint in practice, the AI employed in the study is not trained over the changed problem data to simulate the situation when a real-world design constraint falls outside the training scope of an AI. The AI results in colored heatmap suggestions that must be interpreted by humans

based on color intensity shown on the GUI. The AI is expected to significantly improve human design team performance in the study since prior research has shown that the AI outperforms human designers when working independently (Raina, McComb, & Cagan, 2019). Research findings from the current study inform development of design AIs and the means to incorporate them into design teams effectively. Importantly, this research also provides several warnings for using AI to solve engineering design problems.

This paper begins with a description of the configuration design problem and the procedure of the human subject study in Section 1. The background framework for the deep learning AI and the guidelines to interpret the heatmap suggestions from the AI are presented in Section 2. The results of the human subject study are reported and the impact of the AI on the performance, behavior, and perceived workload of distributed design teams is analyzed in Section 3. In Section 4, the potential reasons behind the results of the human subject study are discussed, and several suggestions and warnings for utilizing AI in engineering design are provided accordingly. The paper concludes with key findings from the human subject study and a brief discussion of the contribution of the research.

1 Human subject study overview

To understand the impact of AI on human design teams, a human subject study is performed. This section describes the configuration design problem and the GUI employed in the human subject study. Information about the pool of participants and the procedure of the study are also presented.

1.1 GUI and problem description

In the human subject study, participants in teams of three are asked to design truss bridges that satisfy specified design requirements using a GUI. The truss bridge design GUI is shown in Figure 1. The GUI allows each participant in the study to add or delete a joint, add or delete a two-force member between two joints, increase or decrease the thickness of a member, move the position of a joint, increase or decrease the thickness of all members, delete all members, and undo or redo previous actions. The GUI always informs participants about whether the current truss bridge design is acceptable and provides factor of safety and mass of the current design based on the predefined loads and supports (Hibbeler, 2012; Rahami, 2007). Participants can also load their teammates' designs through the GUI anytime in the study. Each action participants perform in the study is recorded by the GUI for post-study analysis.

Among the human design teams in the study, half of the design teams are human-only teams with no available deep learning AI. The other half of the design teams, designated as hybrid teams, have a deep learning AI to advise

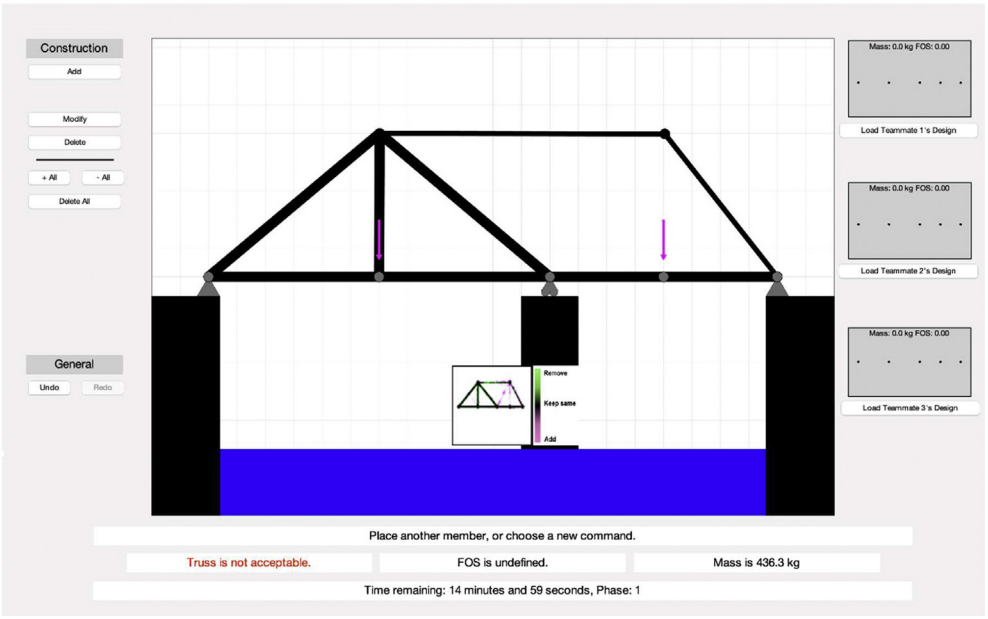

Figure 1 Truss bridge design GUI with AI heatmap suggestions in the first session

them for design modifications to the evolving truss bridge. Suggestions generated by the deep learning AI are provided as a heatmap, shown in Figure 1. Participants in the hybrid teams must interpret the heatmap suggestions based on intensity of color (e.g., pink line suggests adding a member between two joints, and green line suggests deleting a member between two joints). The details about the deep learning AI development and heatmap suggestions interpretation are provided in Section 2.

The human subject study includes two sessions. At the beginning of the first session, the problem statement (PS1) given to all participants is as follows:

1. Design a bridge that spans the river, supports a load at the middle of each span and has a factor of safety greater than 1.25.
2. Achieve a mass that is as low as possible, preferably less than 175 kg.

Each participant then starts to design a truss bridge from scratch in the first session. Figure 1 shows a representative truss bridge design with AI heatmap suggestions based on the problem statement (PS1). Halfway through the study, the design problem abruptly changes. The problem change emulates design specification changes that often take place in practice during a design execution to solve a technical issue, to improve performance of the design, or to adapt to market changes (Sudin & Ahmed, 2009; Zhang, Morris, Allaire, & McAdams, 2020). Here, the abrupt inclusion of an obstacle to be avoided allows for the examination of the responses of the hybrid teams and

the human-only teams to such a problem change. The new problem statement (PS2) given to all participants in the middle of the study is as follows:

1. Design a bridge that spans the river, supports a load at the middle of each span and has a factor of safety greater than 1.25.
2. Ensure that the bridge does not overlap or pass through the orange region.
3. Achieve a mass that is as low as possible, preferably less than 200 kg.

As specified in the new problem statement (PS2), participants need to avoid an obstacle (i.e., the orange region) when they design their truss bridges in the second session. Figure 2 shows a representative truss bridge design with AI heatmap suggestions based on the new problem statement (PS2). The L-shaped obstacle is located in the middle of the design space and therefore most of acceptable truss bridge designs in the first session become unacceptable in the second session. Of note, the deep learning AI is trained by the dataset collected from the first session of a previous human subject study (McComb et al., 2015; McComb, Cagan, & Kotovsky, 2018) that uses the same truss bridge design GUI but the AI training does not include the problem statement change data. As a result, the heatmap shown in Figure 2 suggests adding a member that passes through the orange region (i.e., a horizontal pink line), which leads to an unacceptable design based on the new problem statement (PS2).

1.2 Participants information

Undergraduate and graduate students are recruited from two engineering courses in Carnegie Mellon University as participants of this human subject study. These students have learned the fundamentals of truss design in previous coursework. In total, 72 students are recruited and randomly assigned to 24 teams of three. Among these 24 teams, 12 teams are the hybrid teams, and the other 12 teams are the human-only teams. Each participant receives course credit and $5 cash compensation at the end of the study. Participants in the best performing hybrid team and in the best performing human-only team each are given an extra $10 gift card as a reward after the study, and participants are told in advance that would be the case.

1.3 Procedure of the human subject study

The human subject study takes 44 min. Time allocation of the study appears in Figure 3. Each participant registers and then is assigned to one of two computer labs with a given team number and a computer number. These two computer labs accommodate the hybrid teams and the human-only teams, respectively. However, participants are not aware that the human subject study has two conditions (i.e., with and without the deep learning AI). Participants do not know who their teammates are and do not sit next to their

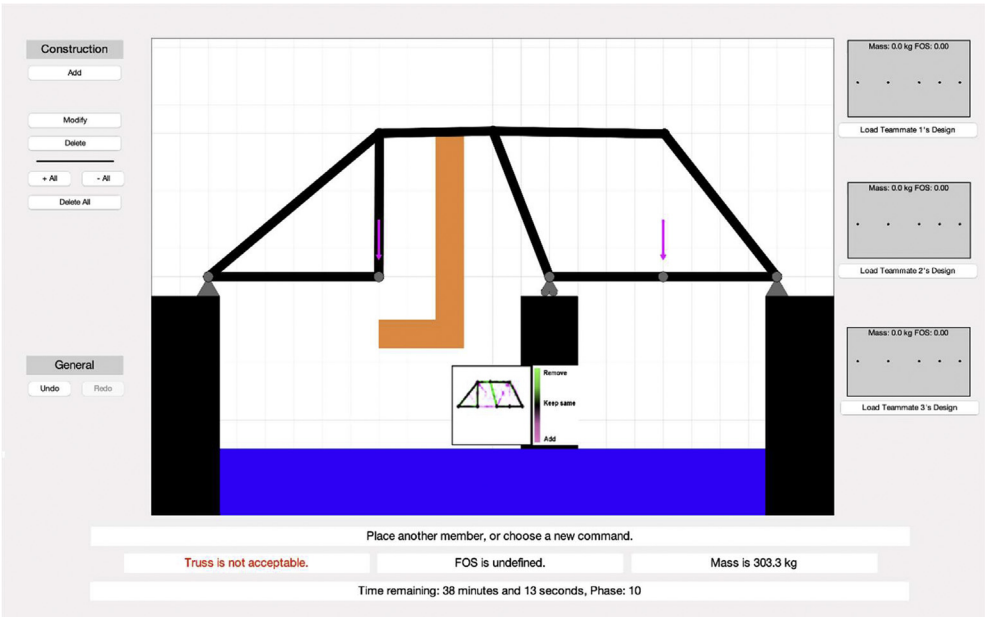

Figure 2 Truss bridge design GUI with AI heatmap suggestions in the second session

	The First Session							The Second Session							
10	2	4	0.5	4	0.5	4	1	2	4	0.5	4	0.5	4	1	2
Tutorial	Read PS1	Team Design	Select Best Design	Team Design	Select Best Design	Team Design	Select Best Design	Read PS2	Team Design	Select Best Design	Team Design	Select Best Design	Team Design	Select Best Design	Post-experiment Questionnaire

Figure 3 Time allocation of the human subject study (numbers indicate duration in minutes)

teammates in the computer lab (mimicking geographically distributed teams). Verbal communication is also not allowed in the study. Participants collaborate with their teammates by sharing their designs through the GUI.

Each participant accesses a separate computer in the lab and logs into the GUI using the given team number and computer number. Participants then complete an interactive tutorial individually. The interactive tutorial walks through each function in the GUI, which provides participants basic information about truss bridge design and team collaboration using the GUI. Participants in the human-only teams have 10 min to complete the tutorial. Participants in the hybrid teams have 7 min to complete the same tutorial

and then are given a 3-min presentation about heatmap suggestions interpretation.

After the 10-min tutorial, the initial problem statement (PS1) stated in Section 1.1 is given to all participants. Participants have 2 min to read the initial problem statement and then begin to design truss bridges from scratch. As shown in Figure 3, the first session includes three 4-min team design periods. Participants design their own truss bridges or load a better or preferred design from their teammates at any time into their own design window for further modification. These three team design periods are divided by two 30-s interludes (i.e., "Select Best Design" with the number of 0.5 shown in Figure 3). These interludes facilitate team collaboration by asking participants to choose whether to continue their own current design or assimilate the best available design created by one of their teammates from which they will work. In other words, each participant has three options (i.e., his or her own current design, the best available design from one teammate, and the best available design from the other teammate) to choose from in each of the 30-s interludes. These interludes require participants in a team to interact, although they are able to do so freely at other times as well. In the end of the first session, each participant has 1 min to select the best truss bridge design from the three best available designs created by themselves and their teammates, respectively. After the 1-min best design selection, participants are given a new problem statement (PS2) as stated in Section 1.1 and begin the second session. The procedure and time allocation of the second session are identical to that of the first session. The only difference is that participants need to avoid the L-shaped obstacle when they design truss bridges based on the new problem statement (PS2) in the second session.

Each participant is asked to fill out a post−experiment questionnaire after they complete the second session. The questionnaire is in paper and pencil and includes seven questions. The first six questions, identical for all participants, are from the official NASA Task Load Index (TLX), which assesses the perceived workload of humans in six scales: mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart & Staveland, 1988). The last question for participants in the hybrid teams quantifies the human confidence in AI in the study. The question is "if the AI heatmap gave you advice 100 times during your design, how many times would you follow the advice given by AI heatmap?", and each participant in the hybrid teams chooses a number from 0 to 100 as the answer. The last question for participants in the human-only teams evaluates their self-confidence in solving the design task. Participants are asked "if you tried 100 times to improve the factor of safety of the 2D truss during your design, how many times would you improve the factor of safety successfully?" Participants in the human-only teams also choose a number from 0 to 100 as the answer of this question.

2 Deep learning AI development and heatmap suggestions interpretation

In the human subject study, a deep learning AI advises participants in the hybrid teams for their truss bridge design. The deep learning AI provides heatmap suggestions in real time, and participants in the hybrid teams must interpret these heatmap suggestions based on intensity of color. This section presents the deep learning framework for the AI development, introduced by a prior research (Raina, McComb, & Cagan, 2019). This prior research has also shown that the AI outperforms human designers when working independently. Importantly, as already mentioned, the deep learning framework is trained by a set of sequential design states images collected from the first session of a previous human subject study that does not include the problem change. Further, the design performance metrics (i.e., factor of safety and mass) are not employed to train the AI. When used by humans, the heatmap suggestions generated by the deep learning AI must be interpreted and the guidelines to do so are provided in Section 2.2.

2.1 Deep learning framework for AI development

This paper seeks to understand how deep learning AIs affect the performance, behavior, and perceived workload of distributed design teams. A deep learning framework constructed and trained by Raina et al. (Raina, McComb, & Cagan, 2019) serves as the AI that aids truss bridge design in the human subject study. The deep learning framework, shown in Figure 4, takes five sequential design state images as the input. Participants only make one action (e.g., add or remove a joint) between two adjacent design states. These images are passed through convolutional *encoder networks* (Glorot, Bordes, & Bengio, 2011; Simonyan & Zisserman, 2014; Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014) that reduce the dimension of each image down to a latent representation while preserving the essential information of the image. A *transition network* (Xu, Wang, Chen, & Li, 2015) then uses the latent representations to predict a representation of the next design state. A *generator network* (Zeiler, Krishnan, Taylor, & Fergus, 2010) maps the predicted representation to the original input size of the image, which is super-imposed with the original image to generate a 76×76 units colored heatmap showing the differences between the before and after images as the output of the deep learning framework.

The deep learning framework including the encoder networks, the transition network, and the generator network, is trained by a set of sequential design state images collected from the previous human subject study that uses the same truss bridge design GUI (but does not include the problem statement change in the data) (McComb et al., 2018). The deep learning AI introduced in this section is also not available to the participants in the previous human

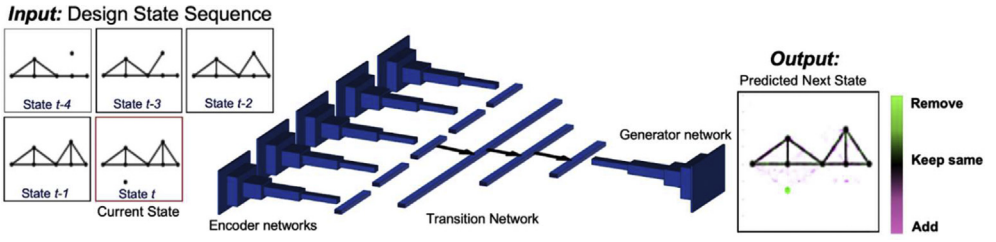

Figure 4 A deep learning framework for AI development

subject study. The training of the deep learning framework includes two steps. In the first step, the encoder network and the generator network are combined to form an autoencoder (Hinton & Salakhutdinov, 2006) where the encoder network reduces the dimension of an input image to a latent representation and the generator network maps the latent representation back to the original image. In the second step, the transition network is trained by sequential latent representations generated by the trained encoder networks. The sequential latent representations follow the chronological order of the original design states and therefore implicitly capture strategies for truss bridge design between these states.

In the two-step training of the deep learning framework, mean squared error is defined as the loss function and Adam optimizer (Kingma & Ba, 2014) is employed to estimate the weights in the autoencoder and the weights in the transition network. An 80/20 split (training vs. testing) is applied to train and test the autoencoder and the transition network. The trained autoencoder and the trained transition network achieve a binary accuracy of 91% and 90% in the test, respectively. Three representative comparisons of the current design state, the colored heatmap generated by the trained deep learning AI, and the actual next design state in the dataset (i.e., ground truth) are shown in Figure 5. In the colored heatmaps, pink lines/points suggest adding members/joints, and green lines/points suggest deleting members/joints. Comprehensive guidelines to interpret heatmap suggestions are provided in Section 2.2. The comparisons in Figure 5 show that the trained deep learning AI is able to predict the actual next design state successfully through the colored heatmaps. Details of the method are found in the prior research (Raina, McComb, & Cagan, 2019).

2.2 Guidelines to interpret heatmap suggestions

Each participant in the hybrid teams will see a new heatmap generated by the deep learning AI every time after the participant is able to take an action, such as add or delete a joint, add or delete a member between two joints, and increase or decrease the thickness of a member. One heatmap usually includes multiple possible suggestions for next actions for the current truss bridge

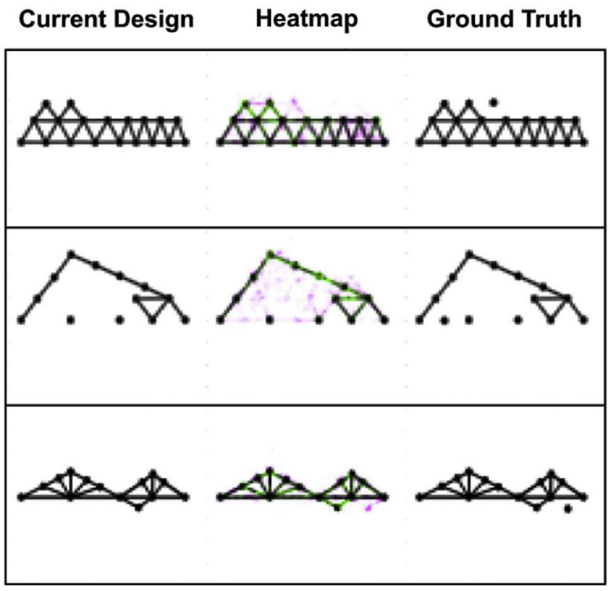

Figure 5 Representative comparisons of the current design, the heatmap, and the ground truth

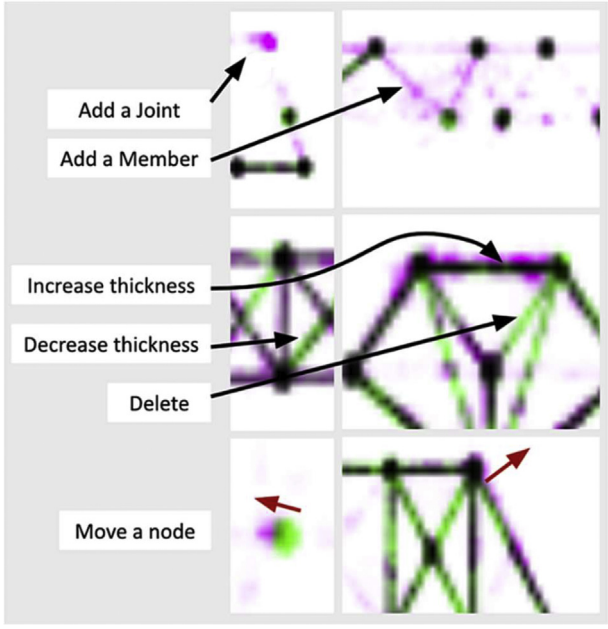

Figure 6 Guidelines for heatmap suggestions interpretation

design. Participants must interpret these suggestions based on intensity of color shown in the heatmap. Figure 6 illustrates several guidelines to interpret heatmap suggestions. These guidelines are listed as follows:

1. If a pink point appears in a blank area, the pink point suggests adding a new joint at the place;
2. If a pink line appears between two joints and no member connects these two joints, the pink line suggests adding a member between these two joints;
3. If a pink line appears between two joints and a member has already connected these two joints, the pink line suggests increasing the thickness of the member;
4. If a green point appears on an existing joint, the green point suggests deleting that joint;
5. If a green line completely covers an existing member between two joints, the green line suggests deleting that member;
6. If a green line appears on an existing member between two joints but does not completely cover the member, the green line suggests decreasing the thickness of the member;
7. If a green point appears on a joint and a pink point appears nearby, these two points suggest moving the joint in the direction from the green point to the pink point.

As stated in Section 1.3, participants in the hybrid teams are given a 3-min presentation about heatmap suggestions interpretation before the first session in the human subject study. The guidelines illustrated in Figure 6 are explained in the 3-min presentation. Figure 6 is also shown on several separate projection screens during the study, so participants in the hybrid teams have free access to Figure 6 anytime in the study.

3 *Results and analysis*

As stated in Section 1, 72 participants are recruited and randomly assigned to 24 distributed teams of three. Among these 24 distributed teams, 12 teams are the hybrid teams, and the other 12 teams are the human-only teams. The human subject study is performed following the procedure defined in Section 1.3. This section presents the results of the human subject study in terms of team performance, behavior, perceived workload, human confidence in AI, and human self-confidence. The results of the hybrid teams are compared with the results of the human-only teams to assess the impact of the deep learning AI on human teams in the study. Further comparisons are made by splitting the 12 hybrid teams and the 12 human-only teams into 4 high-performing teams, 4 middle-performing teams, and 4 low-performing teams, respectively, based on the ultimate team performance in the first session.

Results in Section 3.1 and Section 3.2 are reported using a combination of two-sample t-tests and Cohen's d effect sizes. While a p-value from a statistical test indicates whether a difference between samples is statistically significant, an effect size indicates the magnitude of difference between two samples. An effect

size of 0.2 is generally considered to be small, an effect size of 0.5 is considered to be moderate, and an effect size of 0.8 or larger is considered to be large (Cohen, 1988). Results in Section 3.3 and Section 3.4 are reported using a combination of the Mann−Whitney U tests and Cohen's d effect sizes because ordinal data are collected from the post−experiment questionnaire.

The results of this study demonstrate that the AI boosts the initial performance of low-performing teams in the first session but by the end of the first session there is parity. Once the problem changes, the AI does not help or hurt the performance of low-performing teams. In contrast, the AI hurts the performance of high-performing teams in both sessions. The results also indicate that the AI reduces the number of actions taken by high-performing teams (very large effect size and marginally significant) in the study. In the post−experiment questionnaire, participants in the high-performing hybrid teams believe they accomplish the design task more successfully compared to participants in the high-performing human-only teams, but in fact their performance is worse. Participants in the high-performing hybrid teams also perceive less mental demand compared to participants in the high-performing human-only teams. In addition, most of the participants in the hybrid teams have either high or low confidence in the AI, with few having moderate confidence. In contrast, almost all participants have neither extremely high nor extremely low self-confidence.

3.1 Team performance

Both problem statements (i.e., PS1 and PS2) require participants to maximize factor of safety (FOS) and minimize mass (M) when they design their truss bridges, the strength-to-weight ratio (SWR) is therefore employed to analyze team performance in the study. The strength-to-weight ratio (SWR) is derived by

$$SWR = \frac{FOS}{M}, \tag{1}$$

where mass (M) has the unit of kilogram (kg), factor of safety (FOS) is dimensionless, and thus the strength-to-weight ratio (SWR) has the unit of kilogram^{-1} (kg^{-1}). The best available truss bridge design of each team over the study period is tracked. The average SWR of the 12 hybrid teams is compared with that of the 12 human-only teams to evaluate the impact of the AI on team performance in the study. The comparison of average SWR over the study period appears in Figure 7(a). Figure 7(a) shows that the average SWR of the hybrid teams is lower than that of the human-only teams at the beginning of the first session. The difference in average SWR in the end of the first session and at the beginning of the second session between the hybrid teams and the human-only teams is not significant. In the end of the second session, the average SWR of the hybrid teams is lower than that of

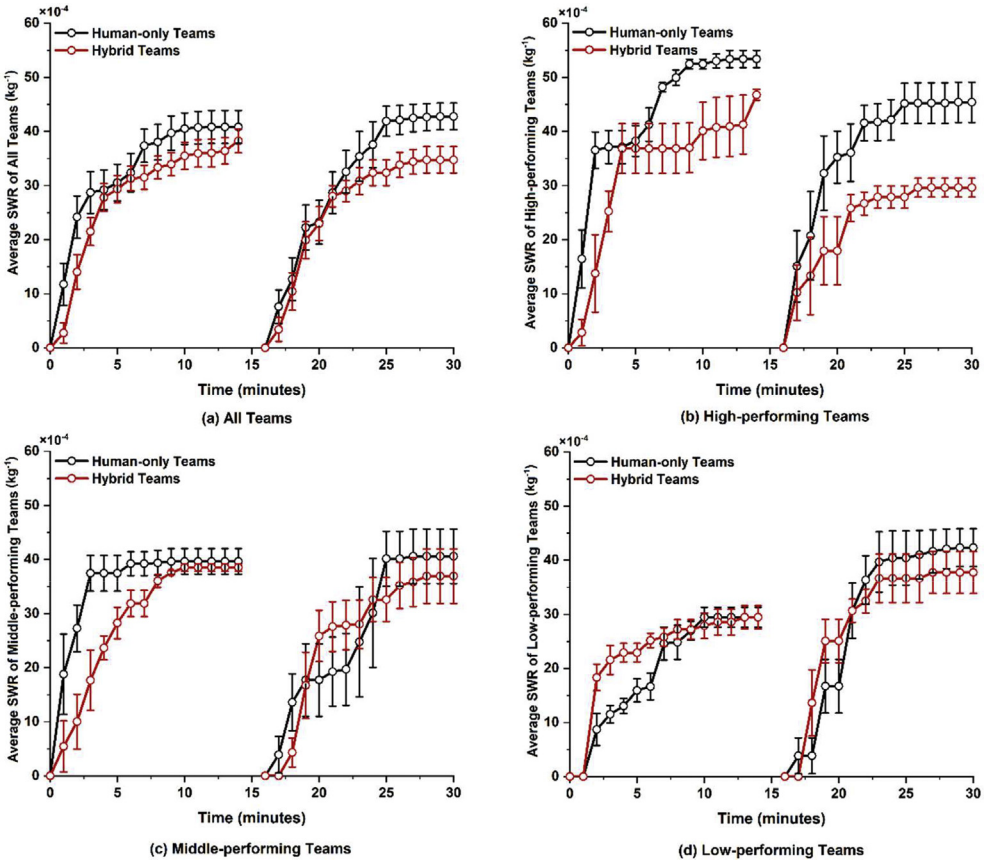

Figure 7 SWR *comparisons between the hybrid teams and the human-only teams over the study period (error bars show ±1 standard error)*

the human-only teams with an effect size of 0.91 and a p-value of 0.037 (the average SWR of the hybrid teams is $34.76 \pm 2.44 \times 10^{-4}$ kg^{-1}, and the average SWR of the human-only teams is $42.77 \pm 2.45 \times 10^{-4}$ kg^{-1}). Of note, several steep slope increases appear in Figure 7(a) (e.g., the performance of human-only teams in the seventh minute) when one or two teams attain significantly better truss bridge designs with the study only including 24 teams in total.

Prior human subject studies show that high-performing teams and low-performing teams have different approaches to solve design problems and overcome abrupt problem changes in the design process (McComb et al., 2015; Ocker & Fjermestad, 2008). Here, to assess the impact of the deep learning AI on distributed design teams with different performance levels, the 12 hybrid teams are divided into 4 high-performing hybrid teams, 4 middle-performing hybrid teams, and 4 low-performing hybrid teams based on the ultimate SWR of each team in the first session. Such division is also applied to the 12 human-only teams. Comparisons in average SWR over the

study period are made between the high/middle/low-performing hybrid teams and the high/middle/low-performing human-only teams, respectively, as shown in Figure 7(b), Figure 7(c), and Figure 7(d). Several steep slope increases also appear in these figures when one or two teams attain significantly better truss bridge designs.

The comparison in average SWR of high-performing teams over the study period appears in Figure 7(b). The average SWR of the high-performing hybrid teams is significantly lower than that of the high-performing human-only teams in the whole study except two short periods when they align. Specifically, the average SWR of the high-performing hybrid teams is lower than that of the high-performing human-only teams in the end of the first session with an effect size of 2.2 and a p-value of 0.023 (the average SWR of the high-performing hybrid teams is $46.77 \pm 1.00 \times 10^{-4}$ kg^{-1}, and the average SWR of the high-performing human-only teams is $53.38 \pm 1.59 \times 10^{-4}$ kg^{-1}); the effect size increases to 2.4 with a p-value of 0.016 in the end of the second session (the average SWR of the high-performing hybrid teams is $29.60 \pm 1.75 \times 10^{-4}$ kg^{-1}, and the average SWR of the high-performing human-only teams is $45.39 \pm 3.71 \times 10^{-4}$ kg^{-1}). These results indicate that the deep learning AI hurts the performance of high-performing teams in both sessions.

The comparison in average SWR of middle-performing teams over the study period appears in Figure 7(c). Figure 7(c) shows that the average SWR of the middle-performing hybrid teams is significantly lower than that of the middle-performing human-only teams at the beginning of the first session. The difference in average SWR in the end of the first session and in the second session between the middle-performing hybrid teams and the middle-performing human-only teams is not significant. These results indicate that the deep learning AI hurts the initial performance of middle-performing teams in the first session. The AI does not help or hurt the performance of middle-performing teams in the subsequent study period.

The comparison in average SWR of low-performing teams over the study period appears in Figure 7(d). As shown in Figure 7(d), the average SWR of the low-performing hybrid teams is significantly higher than that of the low-performing human-only teams at the beginning of the first session. There is no significant difference in average SWR between the low-performing hybrid teams and the low-performing human-only teams in the end of the first session and in the second session. These results suggest that the AI boosts the initial performance of low-performing teams in the first session. The AI does not have significant impact on the performance of low-performing teams in the latter part of the first session. Importantly, although the AI has been trained by the data that does not include the problem change and participants must

interpret the heatmap suggestions from the AI, the AI does not hurt the performance of low-performing teams in the second session.

3.2 Team behavior

The number of actions and the number of collaborations of each team in the study are calculated, respectively. Comparisons are made between the hybrid teams and the human-only teams to assess the impact of the AI on team behavior in the study. As stated in Section 1.1, the truss bridge design GUI records each action (e.g., add a member and add a joint) that participants perform in the study. The total action number of a team is calculated by adding the number of actions from each of the three participants in the team together. The total collaboration number of a team is derived by counting how many times participants in the team load other teammates' designs over the study period. Loading a teammate's design is not counted as a collaboration in the case a participant discards the loaded design immediately.

The comparison in average action number between the hybrid teams and the human-only teams appears in Figure 8(a). Figure 8(a) shows that the difference in average action number between the hybrid teams and the human-only teams is not significant. The average action number of the high/middle/low-performing hybrid teams are also compared with that of the high/middle/low-performing human-only teams, respectively. As shown in Figure 8(a), the average action number of the high-performing hybrid teams is lower than that of the high-performing human-only teams with an effect size of 1.5 and a p-value of 0.078 (the average action number of the high-performing hybrid teams is 700.25 ± 48.52, and the average action number of the high-performing human-only teams is 897.00 ± 63.82). No significant difference is found in the other two comparisons. These results show that the deep learning AI reduces the number of actions of high-performing teams with a very large effect size and marginally significant.

The comparisons in average collaboration number between the hybrid teams and the human-only teams with different performance levels are shown in Figure 8(b). The difference in average collaboration number between the hybrid teams and the human-only teams is not statistically significant in each of these comparisons. These results suggest that the AI does not have significant impact on the collaboration number of human design teams in the study.

3.3 Perceived workload of participants

As stated in Section 1.3, participants fill out a post−experiment questionnaire at the end of the study. The first six questions of the questionnaire are questions in the official NASA Task Load Index (TLX). NASA TLX assesses perceived workload of each participant in six scales: mental demand, physical

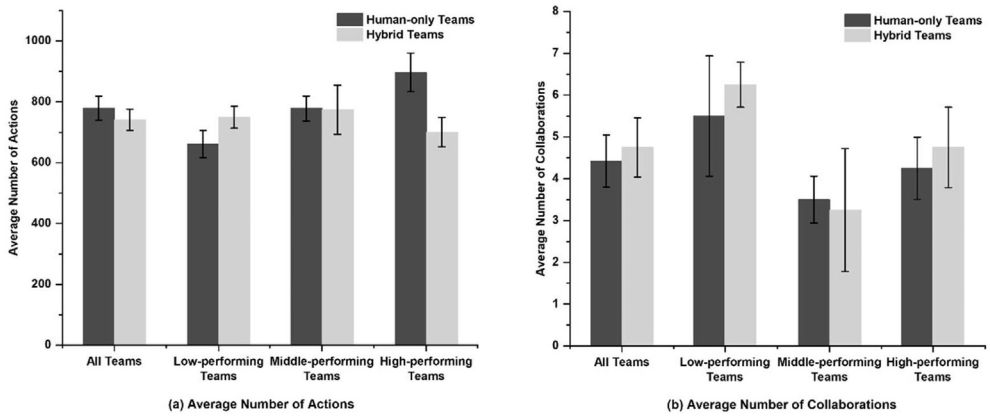

Figure 8 Comparisons in average action number and average collaboration number between the hybrid teams and the human-only teams in the study (error bars show ±1 standard error)

demand, temporal demand, performance, effort, and frustration (Hart & Staveland, 1988). Each question corresponds to one scale. Participants choose a number from 0 to 100 as the answer for each of these six questions. These answers are recorded, and comparisons are made between participants in the hybrid teams and participants in the human-only teams.

The comparisons in each of these six scales between participants in the hybrid teams and participants in the human-only teams appear in Figure 9(a). The figure shows that participants in the hybrid teams perceive less mental demand with an effect size of 0.36, less temporal demand with an effect size of 0.29, and less frustration with an effect size of 0.073 on average than participants in the human-only teams. Participants in the hybrid teams also believe that they accomplish the design task more successfully (with an effect size of 0.16) and spend less effort (with an effect size of 0.14) on average to accomplish their level of performance compared to participants in the human-only teams. However, these results do not have statistical significance ($p > 0.10$ for all results).

The NASA TLX results of participants in the hybrid teams are also compared with that of participants in the human-only teams with three different performance levels respectively, and statistically significant results are only found in the comparison between participants in the high-performing hybrid teams and participants in the high-performing human-only teams, as shown in Figure 9(b). Participants in the high-performing hybrid teams perceive less mental demand on average compared with participants in the high-performing human-only teams with an effect size of 0.96 and a p-value of 0.028. Participants in the high-performing hybrid teams also believe that they accomplish the design task more successfully on average than participants in the high-performing human-only teams with an effect size of 1.0 and a p-

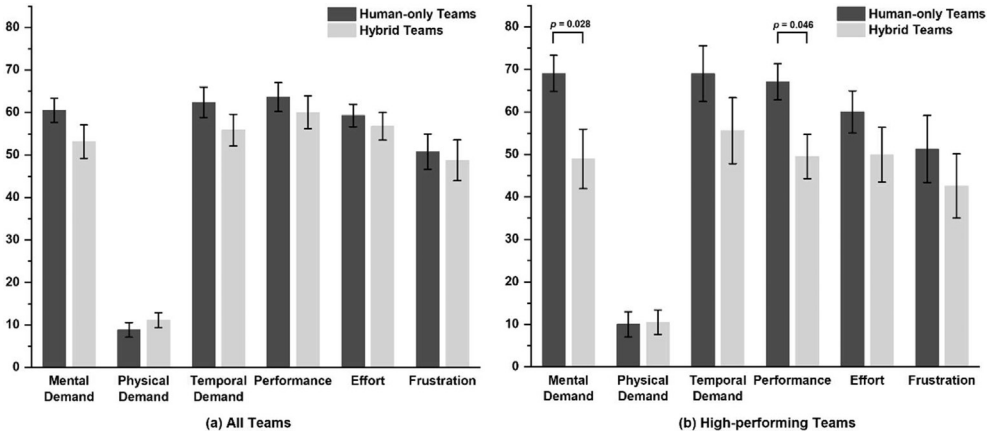

Figure 9 NASA TLX results of all teams and high-performing teams (error bars show ±1 standard error)

value of 0.046. Of note, participants are asked "how successful were you in accomplishing what you were asked to do?" as the question related to performance in NASA TLX. Participants choose a number from 0 to 100 as the answer of that question, where 0 represents "perfect" and 100 represents "failure".

3.4 Human confidence in AI and self-confidence in themselves

As the last question in the post−experiment questionnaire, participants in the hybrid teams are asked "if the AI heatmap gave you advice 100 times during your design, how many times would you follow the advice given by AI heatmap?" to assess the human confidence in AI. Participants choose a number from 0 to 100 as the answer for that question. The distribution of the answers from participants in the hybrid teams appears in Figure 10(a). Figure 10(a) shows that less than 14% of participants have their answers lay between 40 and 60. This result indicates that most of participants in the hybrid teams have either high or low confidence in the AI. The average human confidence in AI between the hybrid teams with different performance levels are also compared. As shown in Figure 10(b), participants have a higher confidence in the AI on average as their teams have a lower performance. The effect sizes of these comparison results are 0.24 (low-performing vs. middle-performing), 0.34 (middle-performing vs. high-performing), and 0.59 (low-performing vs. high-performing), respectively. These comparison results do not have statistical significance, with $p > 0.10$ for all differences in average human confidence in AI between the hybrid teams with different performance levels.

In contrast, participants in the human-only teams are asked "if you tried 100 times to improve the factor of safety of the 2D truss during your design, how many times would you improve the factor of safety successfully?" as the last questions in the post−experiment questionnaire to assess the human self-

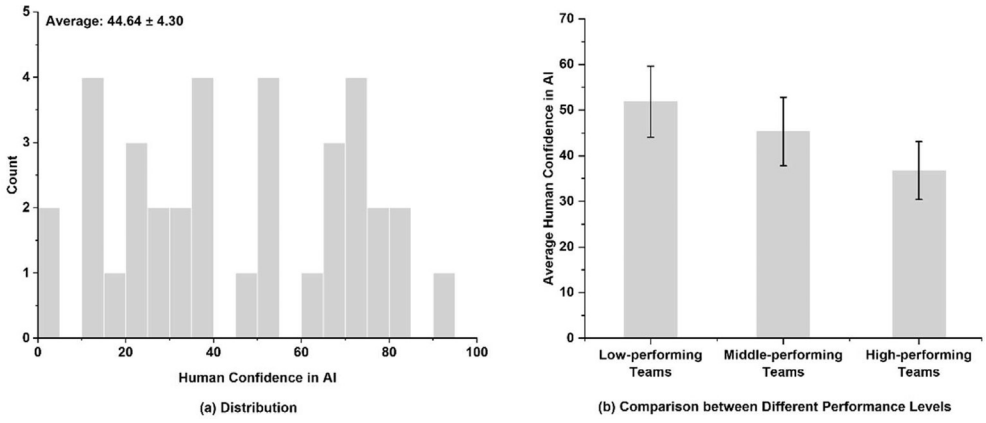

Figure 10 Human confidence in AI results from the post−experiment questionnaire (error bars show ±1 standard error)

confidence. Participants also choose a number from 0 to 100 as the answer for that question. The distribution of the answers from participants in the human-only teams appears in Figure 11(a). As shown in Figure 11(a), less than 3% of participants have their answers larger than 80 or smaller than 20. This result indicates that most participants have neither extremely high nor extremely low self-confidence in the study. The average human self-confidence between the human-only teams with different performance levels are also compared, as shown in Figure 11(b), but no significant result is found.

4 *Discussion*

The results of the human subject study discussed in Section 3 exhibit that the deep learning AI boosts the initial performance of low-performing teams in the first session, but the AI does not help or hurt the performance of low-performing teams by the end of the first session or in the second session (i.e., after the problem statement change). In contrast, the AI hurts the performance of high-performing teams in both sessions. There are three reasons that may explain why the deep learning AI hurts the performance of high-performing teams:

First, heatmap suggestions interpretation takes time and may also cause cognitive overload on participants' visual channel. A heatmap generated by the deep learning AI usually includes multiple design suggestions in the form of color intensity (e.g., pink points/lines and green points/lines). These suggestions are not straightforward for participants to follow. It takes time for participants to interpret these heatmap suggestions based on the guidelines shown in Figure 6 and decide which suggestion given by the AI to adopt or whether to reject all these suggestions. Although participants in the hybrid teams do not report higher perceived temporal demand in the post−experiment questionnaire compared to participants in the human-only teams, participants

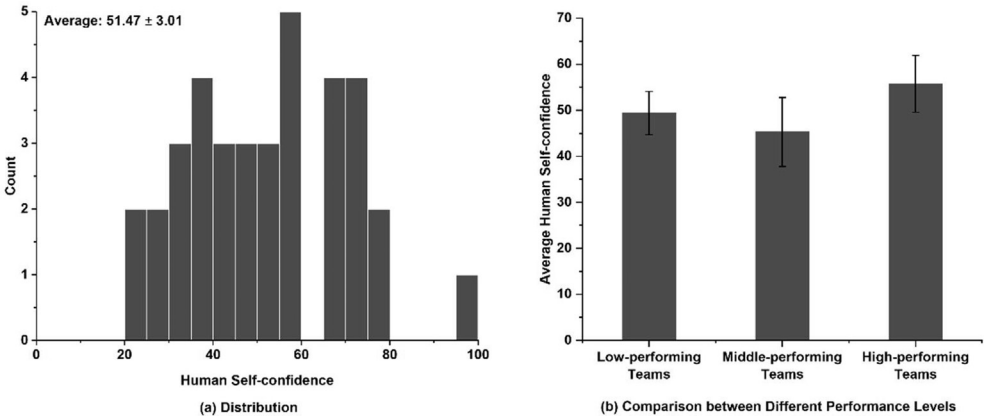

Figure 11 Human self-confidence results from the post−experiment questionnaire (error bars show ±1 standard error)

aided by the AI do have less time to think about the problem itself or steps to solve it. Casual observation indicates that over time this process gets easier, but participants only have 3 min to learn how to interpret the heatmap suggestions. Moreover, the deep learning AI generates a new heatmap for a participant in the hybrid teams every time after the participant takes an action (e.g., add or delete a joint). The human information processing system has an auditory/verbal channel and a visual channel (Paivio, 1990). Participants' visual channel may be overloaded by such successive processing demand in the study since each heatmap includes multiple suggestions that must be interpreted, potentially contributing to general cognitive overload. Cognitive overload, in turn, may reduce information processing efficiency and result in low performance as suggested by several prior psychological studies (Baddeley, 1992; Mayer & Moreno, 2003; Sweller, 1988). Of note, the potential cognitive overload does not lead to higher perceived mental demand of participants in the hybrid teams in the post−experiment questionnaire likely since the availability of the deep learning AI may greatly reduce the perceived mental demand of the participants aided by the AI.

Second, the deep learning AI is trained by sequential design state images created by participants with a whole variety of performance levels in the previous human subject study. In other words, the deep learning AI learns from both high skilled designers and low skilled designers. Despite this, a version of the AI equipped with a rule-based algorithm that automatically makes inferences from the heatmap suggestions has been shown to meet or exceed human solution quality for this problem (Raina, McComb, & Cagan, 2019). Yet it is possible that a participant reaches a design state in the first session, which is close to a low-performing design in the dataset used to train the AI. In that case, although the participant receives multiple suggestions from the heatmap toward high-performing designs, which may require a few more actions to

reach, the participant may not be able to envision these high-performing designs and instead follows a straightforward path of suggestions to attain a low-performing design that the AI was trained on. Moreover, because the AI has not been trained on data from humans avoiding the obstacle, the AI may provide heatmap suggestions that penetrate the obstacle in the second session. Again, the AI independently works around these obstacles, but the participant may take the inferior step. Participants' flawed inference from the heatmap suggestions may give rise to the reduced performance of the high-performing hybrid teams.

Third, the deep learning AI makes participants in the high-performing hybrid teams lazy. The average action number of the high-performing hybrid teams is lower than that of the high-performing human-only teams in the study. The results from the post−experiment questionnaire show that participants in the high-performing hybrid teams perceive lower mental demand and also believe they accomplish the design task more successfully compared to participants in the high-performing human-only teams, but in fact their performance is worse. With the lower mental demand and the illusion of success, participants in the high-performing hybrid teams may be less motivated to create better designs in the study, which may lead to the less human effort in solving the design problem and result in the reduced performance of their teams.

These potential reasons for the reduced performance of high-performing teams in the study provide insights for designers to develop and incorporate AIs into their design teams effectively: (1) the suggestions from an AI should be straightforward for designers to follow or at least easy to interpret; (2) AI also should not provide too many suggestions at a time; and (3) the suggestions from AI should eventually guide human designers towards improved designs.

The results presented in Section 3 and the potential reasons for the reduced performance of high-performing teams also warn designers that AI does not always improve team performance in engineering design. This research shows that the AI used in this work induces an illusion of success for human designers and make them lazy. Once human designers follow AI suggestions, they give up the opportunity to explore the design space by themselves. As shown in the human subject study, design space exploration conducted by human designers, in particular by high skilled human designers, may lead to better performing designs that AI-human collaboration cannot easily achieve.

5 *Conclusions*

The impact of a deep learning AI on distributed design teams is assessed through a human subject study where half of the teams have the deep learning AI to advise them for the design problem, and the other half of the teams work on their own. The design problem abruptly changes in the middle of the study,

adding complexity. The AI boosts the initial performance of low-performing teams only but always hurts the performance of high-performing teams in the study. The reduced performance of high-performing teams is explained through the cognitive overload, the flawed inference from AI suggestions, and a lack of motivation among participants to find better designs in the study. These potential reasons provide insights for researchers to develop AI agents that provide suggestions that are straightforward for human designers to follow or at least easy to interpret, and not overwhelm the user with too many options to consider at a time. The suggestions from AI also should eventually guide human designers towards improved designs.

Although AI has great potential to partner with human problem solvers, it does not always improve performance in engineering design and may result in an illusion of success or reduce human effort in solving a problem. The context and interaction of the AI is critical for effectiveness and must be a core area of focus in the design of effective collaborative AIs.

Declaration of competing interest

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

Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research through grant FA9550-18-0088, and the Defense Advanced Research Projects Agency through cooperative agreement No. N66001-17-1-4064. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

Baddeley, A. (1992). Working memory. *Science, 255*, 556–559.

Camburn, B., Arlitt, R., Anderson, D., Sanaei, R., Raviselam, S., Jensen, D., et al. (2020a). Computer-aided mind map generation via crowdsourcing and machine learning. *Research in Engineering Design* 1–27.

Camburn, B., He, Y., Raviselvam, S., Luo, J., & Wood, K. (2020b). Machine learning-based design concept evaluation. *Journal of Mechanical Design, 142*, 031113.

Chen, H. Q., Honda, T., & Yang, M. C. (2013). Approaches for identifying consumer preferences for the design of technology products: A case study of residential solar panels. *Journal of Mechanical Design, 135*, 061007.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic press.

Dering, M. L., Tucker, C. S., & Kumara, S. (2017). An unsupervised machine learning approach to assessing designer performance during physical prototyping. *Journal of Computing and Information Science in Engineering, 18*, 011002.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 315–323).
Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load Index): Results of empirical and theoretical research. In . *Advances in psychology, Vol. 52* (pp. 139–183). Elsevier.
Hibbeler, R. C. (2012). *Structural analysis*. Upper Saddle River, New Jersey: Pearson Prentice Hall.
Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*, 504–507.
Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.
Lopez, C. E., Miller, S. R., & Tucker, C. S. (2019). Exploring biases between human and machine generated designs. *Journal of Mechanical Design, 141*, 021104.
Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*, 43–52.
McComb, C., Cagan, J., & Kotovsky, K. (2015). Rolling with the punches: An examination of team performance in a design task subject to drastic changes. *Design Studies, 36*, 99–121.
McComb, C., Cagan, J., & Kotovsky, K. (2018). Data on the design of truss structures by teams of engineering students. *Data in brief, 18*, 160–163.
Ocker, R. J., & Fjermestad, J. (2008). Communication differences in virtual design teams: Findings from a multi-method analysis of high and low performing experimental teams. *ACM SIGMIS - Data Base: The DATABASE for Advances in Information Systems, 39*, 51–67.
Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press.
Rahami, H. (2007). Truss analysis. In *Mathworks file exchange software*.
Raina, A., Cagan, J., & McComb, C. (2019). Transferring design strategies from human to computer and across design problems. *Journal of Mechanical Design, 141*, 114501.
Raina, A., McComb, C., & Cagan, J. (2019). Learning to design from humans: Imitating human designers through deep learning. *Journal of Mechanical Design, 141*.
Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556.
Singh, V., & Gu, N. (2012). Towards an integrated generative design framework. *Design Studies, 33*, 185–207.
Song, B., Zurita, N. S., Zhang, G., Stump, G., Balon, C., Miller, S., et al. (2020). Toward hybrid teams: A platform to understand human-computer collaboration during the design of complex engineered systems. In . *Proceedings of the design society: DESIGN conference, Vol. 1* (pp. 1551–1560). Cambridge University Press.
Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). *Striving for simplicity: The all convolutional net*. arXiv preprint arXiv:1412.6806.
Sudin, M. N., & Ahmed, S. (2009). Investigation of change in specifications during a product's lifecycle. In . *DS 58-8: Proceedings of ICED 09, the 17th international conference on engineering design, Vol. 8* (pp. 371–380). Palo Alto, CA, USA: Design Information and Knowledge, 24.-27.08. 2009.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.

Williams, G., Meisel, N. A., Simpson, T. W., & McComb, C. (2019). Design repository effectiveness for 3D convolutional neural networks: Application to additive manufacturing. *Journal of Mechanical Design, 141*.
Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review, 96*, 114–123.
Xu, B., Wang, N., Chen, T., & Li, M. (2015). *Empirical evaluation of rectified activations in convolutional network*. arXiv preprint arXiv:1505.00853.
Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition* (pp. 2528–2535). IEEE.
Zhang, G., Morris, E., Allaire, D., & McAdams, D. A. (2020). Research opportunities and challenges in engineering system evolution. *Journal of Mechanical Design, 142*, 081401.