How does machine advice influence design choice? The effect of error on design decision making

Ananya Nandy and Kosa Goucher-Lambert
University of California, Berkeley
ananyan@berkeley.edu, kosa@berkeley.edu

Engineering design relies on the human ability to make complex decisions, but design activities are increasingly supported by computation. Although computation can help humans make decisions, over- or under-reliance on imperfect models can prevent successful outcomes. To investigate the effects of assistance from a computational agent on decision making, a behavioral experiment was conducted (N = 33). Participants chose between pairs of aircraft brackets while optimizing the design across competing objectives (mass and displacement). Participants received suggestions from a simulated model which suggested correct (i.e., better) and incorrect (i.e., worse) designs based on the global design space. In an uncertain case, both options were approximately equivalent but differed along the objectives. The results indicate that designers do not follow suggestions when the relative design performances are notably different, often underutilizing them to their detriment. However, they follow the suggestions more than expected when the better design choice is less clear.

Introduction

A typical engineering design process consists of several stages, such as planning, concept development, system-level design, detail design, testing and refinement, and production ramp-up, where humans must make many decisions related to the design [1]. They often utilize several methods and sources of information, including objective measures obtained from controlled prototype testing or "rule of thumb" heuristics from experience to make these decisions. Additionally, engineering designers have the increasing ability to augment their design process using computational methods and

Design Computing and Cognition DCC'22. J.S. Gero (ed),
pp. xx-yy. © Springer 2022

models [2]. For instance, generative adversarial networks (GANs) can be used to synthesize general concepts according to text descriptions (e.g., a chair shaped like an avocado) [3], or natural language processing-based machine learning methods can automate the evaluation of early-stage design concepts for novelty [4]. GANs can also be used to synthesize designs that meet specific engineering requirements (e.g., airfoils with particular lift-to-drag ratios [5] or part interdependencies [6]) for the detail design stage. Data-driven surrogate models can then allow engineers to simulate their designs more efficiently and allow design optimization across these numerous design options for the testing and refinement phase [7]. Additionally, intelligent assistants have been developed to augment designers across the various phases of complex system design [8]. While computational models for generation or evaluation of designs can assist decision making, engineering designers typically make the final choices on how to utilize the model outputs for design activities.

The assistance of computational models and interfaces may generally help designers make difficult design decisions or navigate an expansive design space. However, this type of decision making, sometimes referred to as "AI-assisted decision making," relies on the ability to calibrate human trust in the system to ensure that joint decision making leads to an overall improvement in outcomes [9]. Therefore, it is important to investigate what happens when these models are wrong or when they clash with a designer's intuition or intent. For example, how might a human's design decisions be impacted by a concept evaluator that does not find novel concepts or a GAN that generates a design that is not functionally viable? There can be significant financial or safety-related costs to the outcomes (i.e., final designs) if suboptimal decisions are made during the engineering design process. These poor decisions might stem from erroneous judgement on the human side, erroneous information provided by any computational tool that is used, or a combination of both. Engineering designers may have enough domain expertise to recognize errors, at the cost of diminishing trust in the systems.

Engineering design requires making tradeoffs based on a variety of factors, adding another layer of complexity to collaborative interactions. In a case where a design decision may not be clearly "right" or "wrong," how would a model's output be interpreted and inform decisions? There is evidence that in uncertain domains (where uncertainty cannot be resolved until the event has taken place), people, and particularly experts, prefer to use human judgement even when assisted by algorithms that can outperform them [10]. Uncertainty often cannot be resolved until after costly testing or deployment procedures in engineering design, motivating the need to understand how the behavior might persist in this domain. To investigate the behavioral impact of suggestions from a computational model during design, we conducted an experiment where participants were asked to optimize a design while trading off between two objectives. During decisions, participants were provided with the assistance (in the form of a suggested solution) of an imperfect simulated computational model as well as information to make their own judgements. The results provide initial insights into the decision-making behavior and performance of engineering designers in collaboration with an agent during an uncertain, multi-objective task.

Related Work

Human-computer collaboration in engineering design

Human-computer collaboration has been envisioned to take advantage of designers' ability to formalize design problems while overcoming the cognitive limit of the many variables in a design problem in engineering design [2, 11]. Multi-objective optimization algorithms are specifically useful for engineering design problems [12] and prior work finds that bringing humans in the loop for this process, for example, by using a decision-making paradigm called trade space exploration, helps in the search for optimal designs [13]. Similarly, a study of side-by-side human-robot trade space exploration for complex system design finds that collaboration leads to better designs than solo efforts. However, downsides arise when humans become aware of (and sometimes frustrated with) the limitations of the agent and its suggestions [14]. The effect of algorithmic or AI advice has also been examined through tasks such as drone or truss structure design. In the case of drone design, the effect of AI assistance, which provides Pareto optimal design suggestions from a generative algorithm, is measured by its impact on the overall quality (defined as a utility function of several objectives: range, velocity, cost, and payload) of a designer's final design submissions. A between-subjects study reveals that the quality of drone designs is generally higher when participants (self-reported experts and nonexperts) are provided with the AI assistance compared to when they design alone [15]. The truss structure design study investigates the effect of AI assistance on design teams instead of on individuals. Experimental results find that, unlike for the drone design task, the AI assistance appears to hurt the performance (as measured by the strength to weight ratio of the truss) of high-performing teams [16]. These studies demonstrate the budding potential for human-AI collaboration in engineering design by separating participants into conditions where they either have or lack access to the AI assistance and measuring resulting differences in performance [14–16]. However, these performance improvements might only be realized if people are willing to accept AI assistance in the first place. The study conducted here shares similarities with the previous engineering design studies in its task (having multi-objective criteria) but focuses, in addition to resulting performance measures, on the tendency to follow or ignore assistance during the task.

Human trust in automation and acceptance of algorithmic advice

A survey on studies of human interactions with technology reveals that a variety of human factors, such as trust, mental workload, and automation accuracy, affect whether automation (defined in this case as a machine agent carrying out a previously human function) is used by a human or not. Humans can exhibit both overreliance and underutilization of automation, influenced by different combinations of these factors. Overreliance can be caused by using the automation as a decision heuristic, a possibility for experts and nonexperts alike. Underutilization, on the other hand, is often a result of a lack of trust from the human side [17]. Prior work typically frames the lack of proper reliance on "machine advice" as two conflicting human biases: algorithmic appreciation and algorithmic aversion. Algorithmic appreciation refers to humans preferring assistance from an algorithm over another human [18], while algorithmic aversion refers to resistance to accepting recommendations from algorithms (even if they may outperform humans) [19]. Experimental results relating to algorithmic appreciation vs. algorithmic aversion are inconsistent. Several forecasting experiments indicate that people were generally more likely to accept advice from an algorithm than from other humans, lending support for algorithmic appreciation. However, prior work has found that appreciation of algorithmic advice reduces when people choose between the algorithm and their own judgement and when they have domain expertise. Notably, in the forecasting experiments, experts exhibited reduced accuracy compared to nonexperts due to their discounting of the algorithmic advice [18]. Supporting algorithmic aversion, some experiments find that people were likely to disregard suggestions after observing a mistake, even if the algorithmic results outperformed human decisions on average [19]. Experimental data from a perceptual decision-making task also indicates algorithmic aversion behavior, explaining this behavior through a meta-cognitive bandit model [20]. Unlike the tasks in many of these studies, which have a ground truth for comparison, the tasks in the engineering design studies tend to be more open-ended. The study conducted here introduces some of the open-endedness typical of design but maintains similar structure to previous studies on algorithmic appreciation and aversion by utilizing repeated decision-making trials.

Methods

A human subject experiment was conducted to explore the impact of computational model output on design decision making. The study consisted of a task, where participants made repeated decisions, and a post-task survey, where participants answered questions related to the task and themselves.

Participants: Data was collected from 33 participants, recruited from a university's design, mechanical engineering, and materials science engineering departments. Data collection was IRB-approved and participants were compensated $10 for their time. To be eligible for the study, participants indicated that they were over the age of 18 and had taken a course in structural/solid mechanics. Participants ranged from 19 to 31 years old (M = 22.5, SD = 3.0). 21 participants were men, 10 were women, and 2 were non-binary. There were 17 undergraduate students, 13 graduate students (PhD and masters), 2 working professionals, and 1 recent graduate. Finally, 25 participants indicated that they had 0-4 years of engineering/design experience, while 8 indicated that they had 5-9 years of engineering/design experience.

Context: This study utilized the redesign of a jet engine bracket for additive manufacturing. The task was based on the real design challenge hosted by General Electric on GrabCAD, where engineers had to assess a weight vs. strength tradeoff and submit optimized bracket designs, which were then evaluated using simulation [21]. A dataset of these designs, their properties, and simulations of their performance was made publicly available and a subset of this dataset was used as the stimulus set for this study [22]. Participants in our study were provided with pairs of the designs with accompanying information and asked to select the "better" design, utilizing the same tradeoff. The provided information included 3D models of the designs, performance graphs, visualizations of the simulation results, and a suggestion (Fig. 2). Participants were informed that the suggestion was from a computational model trained on many designs (the actual mechanism for determining the suggestion is explained in the next section). This "model" could be erroneous, indicated by whether it correctly suggested a better design according to the multi-objective criteria.

Experimental design: There were 381 bracket designs in the dataset, representing the global design space explored during the GrabCAD challenge. Each design had an associated mass, maximum displacement (determined across the four loading conditions in the original challenge), and category (determined qualitatively [22]). Bracket designs were compared based on Pareto optimality with two equally weighted objectives. The Pareto frontier refers to where no individual criterion can be made better without making another criterion worse. Therefore, to classify the multi-objective performance of each design, the Pareto optimal set was calculated iteratively across the designs as follows: (1) the Pareto optimal set (designs on the Pareto frontier using mass and maximum displacement as criteria) was found across the full set of designs, (2) that optimal set was removed, (3) the Pareto optimal set was calculated again for the rest of the designs, and (4) this process was repeated until each design was in one of the sets. The results of this process are visualized in Fig. 1, where the lower (i.e., earlier) iterations indicate the more globally optimal designs as opposed to the higher (i.e., later) iterations. This method provided a quantification of designs that were "better" or "worse" than each other and those that were similar in performance.

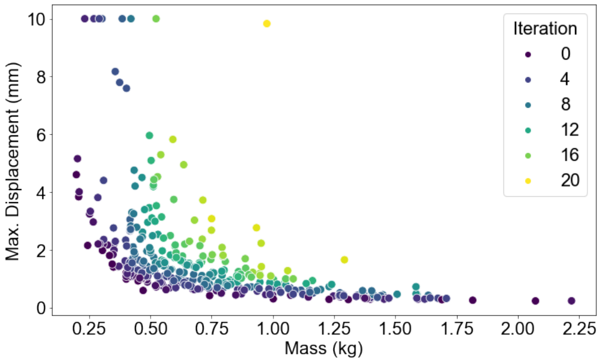

Fig. 1. The Pareto optimal set was iteratively calculated to quantify comparisons between pairs of designs based on multi-objective criteria. Iteration 0 refers to the globally optimal set of designs.

The iteration classifications were used to quantify which design was suggested to the participants as the "model's suggestion." This simulated model emulates a data-driven model in several key aspects. For instance, data-driven models make use of large amounts of data that humans cannot synthesize on their own. In this experiment, while the participants could only access a local set of designs (the two designs they decide between), the "model" assessed which design was better based on the global set (all the designs that were submitted as potential solutions in the challenge). The Pareto optimal sets were calculated based only on designs that were explored during the original challenge, excluding any possibly better designs that were left unexplored and are therefore unknown. Real data-driven models share this limitation, as they may struggle reach new areas of a design space.

The experiment had one manipulation (the model's suggestion—Table 1) and participants were exposed to all conditions (within-subjects) with the trials shuffled pseudo-randomly. The accuracy of the model was set to 71% (only counting trials where there is a “ground truth” for the better design) to try to ensure that the participants did not immediately lose trust in the suggestions. This meant that fewer trials were presented for the incorrect suggestion condition compared to the other conditions.

Table 1. Conditions and trials in the experiment

Condition	**Number of Trials**	**%**
Correct suggestion (suggested design was optimal in an earlier iteration)	15	50
Equivalent suggestion (both designs were considered optimal in the same iteration)	9	30
Incorrect suggestion (suggested design was optimal in a later iteration)	6	20

Constraints on the task length and the number of suitable stimuli limited the number of trials conducted for each participant. Additionally, because participants were required to have some amount of domain expertise within design, the availability of participants was limited.

Stimuli: The stimulus set was selected so that their multi-objective properties would represent the global space (shown in Fig. 1) of designs. A total of 60 bracket designs (two designs per trial for 30 trials) were shown to participants for the main block, with an additional eight for four practice trials and one for the instructions. None were repeated. Several factors were counterbalanced within conditions. First, the distance between the maximum displacement values were balanced since they are found via simulation and may represent a source of error. Next, categories (flat, block, butterfly, beam, and arch), referring to the rough general shapes of the brackets, were balanced as they have notable visual differences. Finally, the locations of the designs in the multi-objective design space were balanced. Some factors were not fully accounted for due to the visual diversity of the designs in the dataset, the qualitative nature of the categorization, and the imbalance of the number of designs in each of these categories. We did not counterbalance if the same underlying stimuli were associated with "correct" and "incorrect" model responses (this would allow for counterbalancing only a subset of the items due to the unequal number of trials in each condition).

Interface: The task was deployed online and participants were directed to a Google Forms survey after completion. The data from the task was collected through a custom online interface developed in Unity (using the UXF [23]) and sent to a database in Amazon Web Services. Figure 2 shows the interface and the types of information available, including an interactable 3D object. Since the two design alternatives were presented side-by-side, a counterbalancing factor was included to account for if the suggested design was on the right or left. Every participant had the same four practice trials, which contained the correct and incorrect suggestion conditions with similar accuracy to the remaining trials (75%). Participants were given feedback on how many times they selected the optimal design during practice (e.g., ¾ times), but no information about which trials they answered correctly.

Data was collected about the designs selected by the participants as well as the time spent on each trial. Prior to analyzing the data, four trials (of 990 trials total, not including practice trials) were removed because the response time was not greater than 500 milliseconds (approximately the time needed to consciously recognize and respond to a visual stimulus).

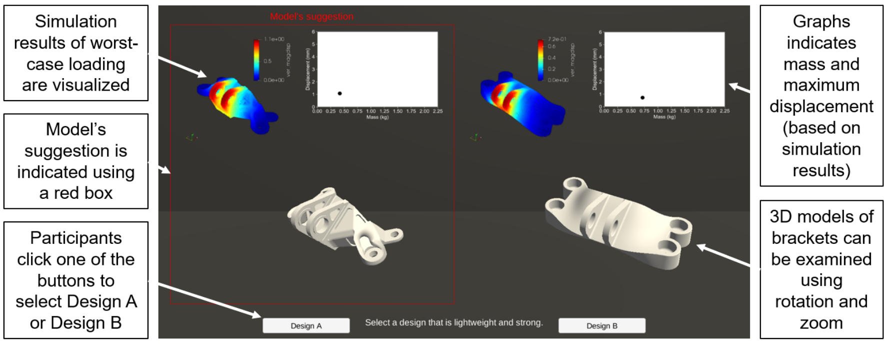
Simulation results of worst-case loading are visualized
Model's suggestion is indicated using a red box
Participants click one of the buttons to select Design A or Design B
Model's suggestion
Displacement (mm)
Mass (kg)
Design A
Select a design that is lightweight and strong.
Design B
Graphs indicates mass and maximum displacement (based on simulation results)
3D models of brackets can be examined using rotation and zoom

Fig. 2. Participants were instructed to select designs based on the information provided, using the interface shown.

Survey: Participants were directed to a survey after the task and asked the questions about the perceived accuracy of the model, the information they used to make their decisions, the strategy they used during the task, their knowledge about the task domain, and their experience in engineering and design more generally.

Results

The effect of model error on suggestion acceptance and performance

The trials were analyzed to investigate how the simulated model's suggestion of a better or worse-performing design solution impacted participants' decision making. Only conditions with a ground truth ("correct suggestion" and "incorrect suggestion," not "equivalent suggestion") were included in these analyses. Figure 3 shows the proportion of participants' decisions that aligned with the provided suggestions (median = 0.57).

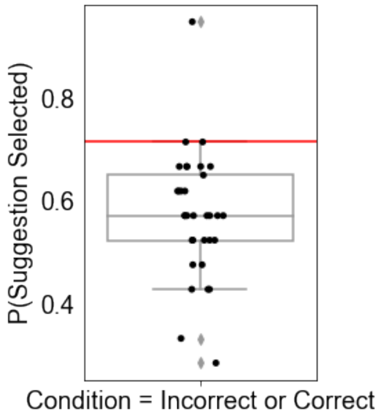

Fig. 3. Distribution of the probability of selecting the suggested design across all participants (median = 0.57) vs. the expected proportion (0.71)

The proportion of decisions that was expected to align with the suggestions – if the participant and the simulated model agreed – was 71%, the actual proportion of correct suggestions. A non-parametric Wilcoxon signed-rank test was conducted as the data violated the assumption of normality, showing that there was a significant difference between the hypothesized value and the observations (W = 24.0, p < 0.001). *When the performance of the two alternatives differed, participants' selection of the suggested design was lower than expected.* The participants' performance was also quantified by their accuracy, referring to the proportion of the time the correct, better-performing design was selected (the same design as the model's suggestion in the correct condition and the non-suggested design in the incorrect condition). Figure 4A shows the accuracy distribution across participants (median = 0.76). Figure 4B displays the accuracy of participants across each of the two conditions. Notably, *participant accuracy was higher when given the incorrect suggestion (median = 0.83) than when given the correct suggestion (median = 0.73).* A non-parametric Wilcoxon signed-rank test was conducted to test this effect, showing that there was a significant difference in the participant accuracy across the conditions (W = 152.0, p = 0.02).

The effect of suggestions in uncertain scenarios

The equivalent condition was analyzed to investigate how the suggestions

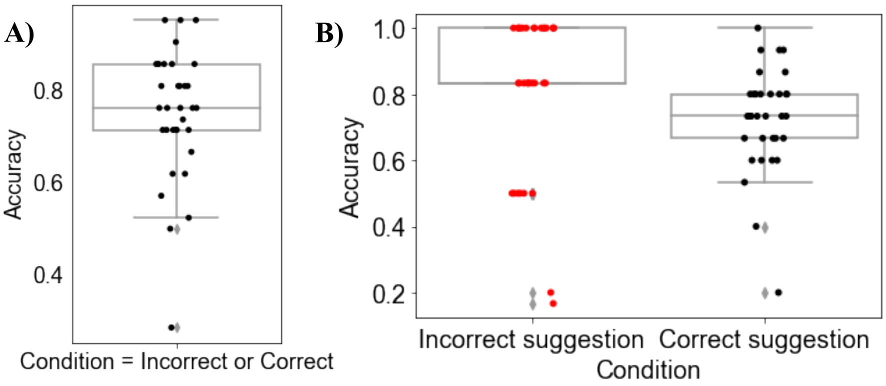

Fig 4. A) Accuracy across all participants (median = 0.76). **B)** Accuracy for incorrect (median = 0.83) vs. correct (median = 0.73) trials (p = 0.02)

impacted participants' decision making when both design alternatives were close in multi-objective performance. These design pairs were considered optimal on the same iteration, but differed in which property was prioritized (e.g., a bracket with high mass but low displacement vs. a bracket with low mass but high displacement). Figure 5A shows a distribution of the proportion of selections that aligned with the model's suggestions across participants (median = 0.78). Considering that the suggestion for this condition was arbitrary, it was expected that the participants' design selections would align with the model's suggestions ~50% of the time. However, a Wilcoxon

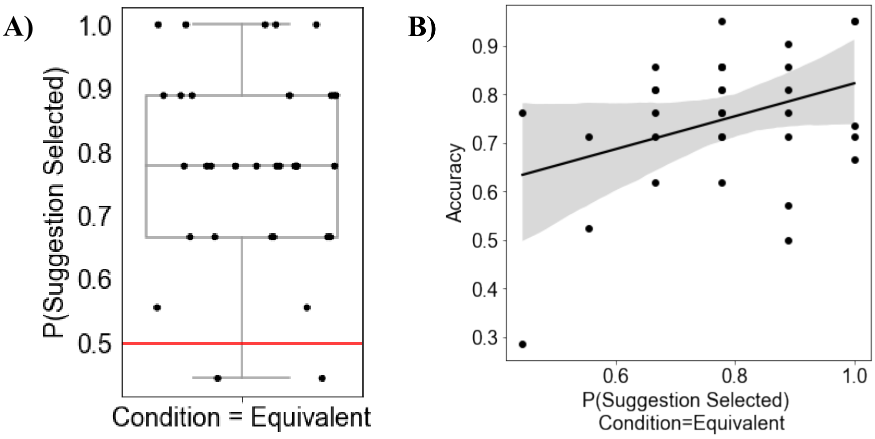

Fig. 5. A) Proportion of trials in the equivalent condition where participants chose what the model suggested (median = 0.78) **B)** Relationship between participants' suggestion selection in the equivalent condition and their accuracy. The two have a moderate positive correlation ($r_p = 0.36$, p = 0.04).

signed-rank test shows a significant difference in participants' proportional selection of the suggested design compared to this expected chance (W = 5.0, $p < 0.001$). *When both design options were close in multi-objective performance, participants chose the model's suggestion more frequently than chance alone.* A Pearson correlation indicates that the proportion of a participant's selections that are aligned with the model's suggestion (in the equivalent condition) is moderately positively correlated with the participant's accuracy ($r_p = 0.36$, $p = 0.04$), shown in Fig. 5B. The accuracy tended to be higher for those who decided to follow the model's suggestion even when both choices were valid as the "better" design.

Participant perception of model accuracy and self-reported decision-making strategies

The effect of several self-reported factors related to knowledge and perceptions were analyzed to determine if these factors were related to participants' performance and decision making. There was no significant Spearman correlation between accuracy and participants' self-reported knowledge (1-7 Likert scale value) in the topics of structural mechanics ($r_s = 0.18$, $p = 0.32$) and multi-objective optimization ($r_s = 0.26$, $p = 0.88$). Looking more closely at the self-reported perceived model accuracy, Fig. 6 shows the distribution of the perceived accuracy has a median of 70%.

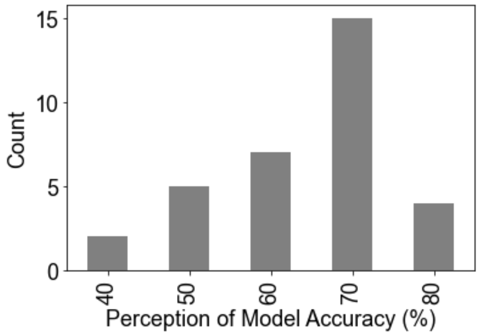

Fig. 6. Distribution of participants' answer to the question "What percentage of the time do you think the model suggested the better design?" (median = 70%). The answer choices ranged from 0 – 100% in increments of 10.

As the accuracy of the suggestions was set to be 71%, not including the trials in the equivalent condition, it appears that participants were relatively good at assessing how often they were receiving correct/incorrect suggestions.

While participants' design selections revealed the outcomes of their decision making, the answer to survey questions provided more insight into participants' decision-making process. These insights are valuable because they can point towards why participants may have performed poorly or well, in addition to why they may or may not have adhered to suggestions. Answers from the open-ended question ("Please describe your decision- making process and strategy in detail.") explained participants' selections. The highest accuracy, achieved by three participants, was 95%. One of these participants had the following strategy:

> *"...An example thoughts process is: 'If part A has less than half the weight of part B and also less than twice the deflection, then part A has a better strength to weight ratio and is the better part'. If I was given a more explicit cost/benefit function then I could have optimized better, eg 'I need the deflection to be below this value, and from there the lowest weight it best'. If the points were too close or I wasn't sure from the graph, then I may go with the Model's suggestion, since the Model has access to the numerical data and can make a more precise calculation than I can in my head ..."*

Notably, the participant made decisions using their own judgement and synthesis of information but recognized when they would benefit from assistance from the suggestion.

Those who had the lowest accuracy commonly exhibited a preferential weighting of one criterion over the other. For example, one participant (accuracy = 29%) mentioned that *"if the graph was more shallow for one design but the stress was significantly larger, then I would choose the design with the smaller stress,"* indicating that they focused more on the strength criterion vs. the weight criterion. Similarly, some participants focused on the geometry of the bracket, but only considered the geometry with respect to the displacement and not the mass. One participant (accuracy = 50%) noted that *"[they] focus on the 3D model essentially and check if there is some panel with small thickness, where the stiffness would be weak,"* while another (accuracy = 57%) also focused on *"how detailed the model is designed and the edges; whether it is too thick or too thin."* Finally, the decision-making strategy from one participant (accuracy = 52%) provided a detailed view into a preference towards a specific type of bracket, as they wrote *"... I look at the stress distribution over the model, personally, I like to pick the brackets that have a more even stress distribution instead of it just concentrated at the connection collar."* This response specifically illustrates how decisions in design can be influenced by the designer's intuition, which may be at odds with computational outputs (the simulated model here only considered the maximum deformation not the distribution).

Examining qualitatively how participants ranked the information that they used to make their selections in the survey shows that the graph with the objective values is most selected as the most important source of information (21 of 33 participants). No one who reached the highest accuracy among participants ranked the 3D model as the most important information for their decisions. In comparison, a few participants who achieved a low accuracy rank the 3D model as the most important information for their decisions and describe using judgements based on its properties. Across participants, regardless of accuracy, 15 of the 33 participants rank the model's suggestions as the least important source of information, while only one ranks them as the most important source.

Discussion

As computation is used to assist increasingly complex decision making, it is important to understand the effects of erroneous or questionable model output on these decisions. In this study, we examined these effects in a multi-objective engineering design decision context.

Designers utilize agent suggestions less than expected when there is a "ground truth"

We found that participants appeared to rely on their own judgement and not on the suggestions in the incorrect and correct conditions, often selecting the alternative design even in when the correct one (with respect to what was determined as correct in this study) was suggested. A statistically significant difference in participants' selection of suggested designs and the expected proportion (57% instead of 71%) in these conditions indicates suggestion underutilization. This is supported by the self-reported rankings of importance of information sources, where participants tended to rank the model's suggestions lower and other sources higher. However, when the participants were not able to synthesize the provided information to determine the right ratio of mass and deformation, not utilizing the suggestions harmed their performance. The statistically significant higher participant accuracy for the condition where they were given the incorrect suggestion compared to when they were given the correct suggestion (83% vs. 73%) also indicates that participants tended to catch when they were given the incorrect suggestion, but erroneously ignore correct suggestions.

These results align with prior work that observes algorithmic aversion [19, 20] and particularly with the finding that experts hurt their accuracy by disregarding the algorithm's suggestions [18]. The group of participants in our study were not necessarily experts in the topic. Though they were required to meet a minimum amount of knowledge in the general area, we did not assess how participants perform without suggestions. However, relatively high participant accuracy (especially in avoiding incorrect suggestions) and qualitative survey responses indicate at least a baseline level of knowledge, which may explain the similarity in findings. Prior studies of AI-assistance in engineering design have found that AI-assistance improves design quality for solo designers [14, 15], yet in this study, these opportunities for improvement are underutilized. In the context of human-computer collaboration in engineering design, the incorrect and correct conditions represent scenarios where the computer can easily find a "better" design (by ensuring the design is closer to Pareto optimality) compared to a human. Therefore, our finding of algorithmic aversion may be an issue, as it conflicts with humans' desired ability to leverage the strengths of computation.

Delving into the qualitative insights from participants' open-ended answers regarding their strategies demonstrated that low accuracy was sometimes explained by a difference in how the participants were making decisions (weighing one objective over the other) and how the model was providing suggestions (weighing both objectives equally). Thus, low accuracy does not necessarily reflect poor performance or lack of knowledge but can alternatively be a result of the mismatch between what is deemed important for the task. Additionally, these responses reflect realistic situations where a designer/engineer may have to prioritize a specific criterion, although the participants in this task were not instructed to prioritize one over the other. In a more realistic setting, there would likely be more factors involved in the process of selecting a design. For example, the background information for the task indicated that the brackets would be made using additive manufacturing. However, if this information was excluded, the choice of design might be influenced by a participant prioritizing manufacturability, which was not included into consideration for the suggestions. For instance, people with more experience might be more likely to consider the manufacturing of the bracket despite not being explicitly instructed to do so. Though expertise was not explicitly examined in this study, it is possible that differences in expertise across participants would be important with the additional consideration of different manufacturing methods.

Designers are more willing to accept agent suggestions when the "better" design alternative is uncertain

The results indicate that participants made their own informed decisions rather than relying on suggestions in the "ground truth" conditions, even though the "better" design can be determined computationally in these conditions. On the other hand, the equivalent condition represents a collaborative scenario where the strengths of human decision making might be particularly important. A computer can output several viable solutions based on the multi-objective criteria but may not be able to distinguish between them. Counterintuitive to this, participants appeared to readily follow suggestions in the equivalent condition. This was demonstrated by a statistically significantly higher proportion of suggestion selection than expected in the equivalent condition (78% instead of 50%) and qualitative data on decision-making strategy.

Outside of the engineering design domain, a study of AI advice acceptance reveals that when people lose self-confidence, they may begin to rely on poor AI suggestions [24]. While participants did not generally follow model suggestions that were clearly wrong, they may have been more likely to follow arbitrary ones when the decision was less clear. Prior studies also indicate that once a design team starts following the advice, they often stop exploring the design space themselves [16] and that a collaborative agent can decrease the coverage of the design space explored [14]. Therefore, based on the results of this study, it is possible that the concerns above could be raised around the influence of computational outputs on design decisions, even if they are not clearly "poor suggestions."

On the other hand, the correlation between a participant's suggestion selection (equivalent condition) and their accuracy (all other conditions) indicates that participants who were more willing to accept the model's suggestion did better. It is possible that participants were able to figure out when the suggestion was bad and make their own decision. However, those who were less likely to choose the suggestion in the equivalent condition (possibly an indication of less trust in the model) more likely failed to follow the model even when it was correct. This, in turn, affected their accuracy. Thus, the results show the delicate balance necessary to appropriately take advantage of human-computer collaboration in design.

Trial-by-trial design properties and their implications on decision making

To examine if specific properties (i.e., the mass or displacement) of the designs affected suggestion selection in a way that may explain the findings, a follow-up analysis was conducted by considering each trial separately and aggregating across participants. Each trial was therefore examined by condition and the difference in properties between its suggested and non-suggested design. The differences across the correct trials (shown in Fig. 7A) implied that the group (and not just individuals) made decisions with some implicit weighting of criteria. If the suggested design had a higher deformation but the tradeoff for lower mass was not subjectively enough, the suggestion was followed less often (though the mass tradeoff for those trials was considered "enough" by the optimization procedure).

Yet preferential weighting of criteria was not necessarily applied in the equivalent condition. Preferential weighting of the deformation objective might lead participants to select the suggested design only if its deformation was lower in this condition (a negative difference). However, this is not the case. Even in trials where the deformation is higher for the suggested design, most of the participants selected the suggestion, as shown in Fig. 7A. Looking only at the equivalent condition, Fig. 7B shows that lower suggestion selection in this condition does not appear to be dependent on one objective over another. Instead, when the differences between both of the properties were close to zero, the decision was arbitrary and the suggested design was selected by closer to 50% of the participants. The percentage of participants selecting the suggested design was much higher when the designs were similar, but the tradeoff was not as easy to assess (larger differences in mass and displacement). The findings from this analysis provide further evidence that the results likely reflect the impact of the suggestions as opposed to other decision-making behavior.

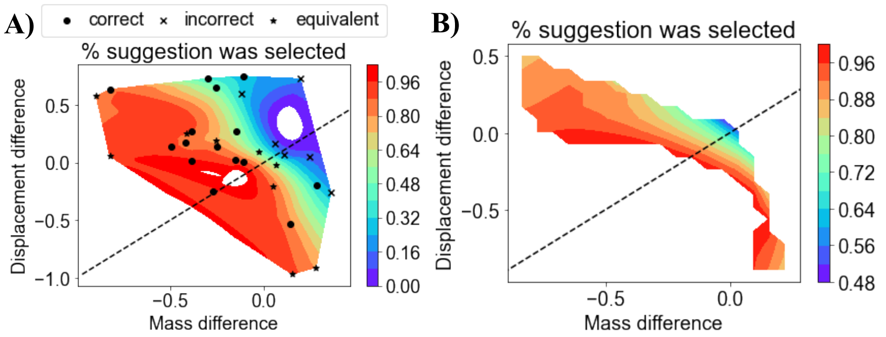

Fig. 7. The difference in stimuli properties along each objective (a positive value indicates a higher value for the suggested design) **A)** for all conditions **B)** as a visualized interpolation for the "equivalent" condition only

Limitations

There are several limitations of this study that should be considered. Design decisions take place over a longer timescale and have more context involved, making it unlikely that the two criteria would have equal weighting, as implied in this study. Additionally, errors in data-driven models related to engineering may be more subtle than the error introduced here, which involved suggesting the wrong design in its entirety. Some limitations relate to the study setup, such as a small sample size and counterbalancing. Though the bracket design pairs presented as stimuli were counterbalanced across several properties (e.g., category, criterion values), there may have been differences in trial difficulty across conditions caused by latent differences in the stimuli sets. This was due to the unequal number of trials across condition (fewer trials for the incorrect condition) and the stimuli subsets never being switched to other conditions. For example, the correct suggestion condition contained a few designs that had low mass but high deformation.

Future Directions

The results from this study point towards many possible avenues for future research. For instance, it is unclear whether the observed higher reliance on arbitrary suggestions would scale with additional complexity in the decision. To address this, it would be necessary to determine how suggestions related to certain characteristics of the design problem might invoke higher or lower levels of trust or acceptance. Studies have revealed that providing people with just the right amount of information can improve trust [25] and perceptions of a model [26]. Including more information about why a design was suggested by the model could impact participants' willingness to accept the suggestions. However, explainability alone cannot address scenarios where a model's "correct" outputs conflict with human decision-making. Instead, adaptivity may be necessary. Recent work has illuminated the importance of mental models [27] and compatibility [28] in human-AI collaboration, offering ways to address these challenges. While the current study does not incorporate these considerations, the results indicate the tendency towards algorithmic aversion, supporting the necessity to account for these factors when developing methods of human-computer collaboration for engineering design. Further examination of decision-making scenarios where a model output could be right or wrong along various dimensions, depending on the human's intent, could be useful for design. For instance, to allow the designer to properly assess how much they should rely on computational systems, it may be necessary for systems to “understand" how the human designer is approaching a problem and communicate if there are differences. Alternatively, a mismatch in decision making might be used to drive automatic adaptation of a computational system to a designer. Uncovering engineering designers' decision-making behaviors in settings where they utilize computational systems can help reveal where general findings around human-AI collaboration apply, and when special considerations must be made for a complex design context. Consequently, this knowledge can be used to develop intelligent tools that effectively fit into human design processes.

Conclusion

The effect of suggestions from a simulated computational model during a design decision-making task is investigated in this study. A jet engine bracket design problem, with a tradeoff between strength and weight, is used to find participants' accuracy in determining the "better" design provided these suggestions. The results indicate that designers' tendency to follow the model's suggestions varies according to the scenario. Designers underutilize suggestions in scenarios where there is a "ground truth," correctly ignoring bad suggestions but also ignoring good ones in the process. This finding might be explained by participants' likeliness to trust their own decision making, even at the risk of performing worse, matching experimental results in other domains that indicate underutilization of algorithmic assistance by those who demonstrate some expertise in the domain. However, when presented with a more uncertain choice between designs, participants tend to follow the model's suggestions more than expected. The results collectively demonstrate the types of behavior that must be accounted for to pursue seamless human-computer collaboration in engineering design.

Acknowledgements

This study was funded by Sandia National Laboratories (SNL). Thank you to Kristin M. Divis, Breannan C. Howell, Laura E. Matzen, and Zoe Gastelum (SNL) for help with experimental design and Whalen et al. for making the bracket data [22] publicly available.

References

[1] K. T. Ulrich and S. D. Eppinger, *Product Design and Development*. McGraw-Hill/Irwin, 2004.

[2] P. Egan and J. Cagan, "Human and Computational Approaches for Design Problem-Solving," in *Experimental Design Research*, P. Cash, T. Stanković, and M. Štorga, Eds. Cham: Springer International Publishing, 2016, pp. 187–205. doi: 10.1007/978-3-319-33781-4_11.

[3] "DALL·E: Creating Images from Text," *OpenAI*, Jan. 05, 2021. https://openai.com/blog/dall-e/ (accessed Sep. 07, 2021).

[4] B. Camburn, Y. He, S. Raviselvam, J. Luo, and K. Wood, "Machine Learning-Based Design Concept Evaluation," *Journal of Mechanical Design*, vol. 142, no. 3, Jan. 2020, doi: 10.1115/1.4045126.

[5] W. Chen and F. Ahmed, "PaDGAN: Learning to Generate High-Quality Novel Designs," *Journal of Mechanical Design*, vol. 143, no. 3, Nov. 2020, doi: 10.1115/1.4048626.

[6] W. Chen and M. Fuge, “Synthesizing Designs With Interpart Dependencies Using Hierarchical Generative Adversarial Networks,” *Journal of Mechanical Design*, vol. 141, no. 11, Sep. 2019, doi: 10.1115/1.4044076.

[7] G. G. Wang and S. Shan, “Review of Metamodeling Techniques in Support of Engineering Design Optimization,” *Journal of Mechanical Design*, vol. 129, no. 4, pp. 370–380, May 2006, doi: 10.1115/1.2429697.

[8] H. Bang, A. V. Martin, A. Prat, and D. Selva, “Daphne: An Intelligent Assistant for Architecting Earth Observing Satellite Systems,” in *2018 AIAA Information Systems-AIAA Infotech @ Aerospace*, American Institute of Aeronautics and Astronautics. doi: 10.2514/6.2018-1366.

[9] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2020, pp. 295–305. doi: 10.1145/3351095.3372852.

[10] B. J. Dietvorst and S. Bharti, “People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error,” *Psychol Sci*, vol. 31, no. 10, pp. 1302–1314, Oct. 2020, doi: 10.1177/0956797620948841.

[11] A. Viros-i-Martin and D. Selva, “A Framework to Study Human-AI Collaborative Design Space Exploration,” in *ASME 2021 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Nov. 2021. doi: 10.1115/DETC2021-67619.

[12] K. Deb and K. Deb, “Multi-objective Optimization,” in *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*, E. K. Burke and G. Kendall, Eds. Boston, MA: Springer US, 2014, pp. 403–449. doi: 10.1007/978-1-4614-6940-7_15.

[13] T. W. Simpson, D. Carlsen, M. Malone, and J. Kollat, “Trade Space Exploration: Assessing the Benefits of Putting Designers ‘Back-in-the-Loop’ during Engineering Optimization,” in *Human-in-the-Loop Simulations: Methods and Practice*, L. Rothrock and S. Narayanan, Eds. London: Springer, 2011, pp. 131–152. doi: 10.1007/978-0-85729-883-6_7.

[14] M. V. Law, N. Dhawan, H. Bang, S.-Y. Yoon, D. Selva, and G. Hoffman, “Side-by-Side Human–Computer Design Using a Tangible User Interface,” in *Design Computing and Cognition ’18*, J. S. Gero, Ed. Cham: Springer International Publishing, 2019, pp. 155–173. doi: 10.1007/978-3-030-05363-5_9.

[15] B. Song, N. F. Soria Zurita, H. Nolte, H. Singh, J. Cagan, and C. McComb, “When faced with increasing complexity: The effectiveness of AI assistance for drone design,” *Journal of Mechanical Design*, pp. 1–38, Jul. 2021, doi: 10.1115/1.4051871.

[16] G. Zhang, A. Raina, J. Cagan, and C. McComb, “A cautionary tale about the impact of AI on human design teams,” *Design Studies*, vol. 72, p. 100990, Jan. 2021, doi: 10.1016/j.destud.2021.100990.

[17] R. Parasuraman and V. Riley, “Humans and Automation: Use, Misuse, Disuse, Abuse,” *Hum Factors*, vol. 39, no. 2, pp. 230–253, Jun. 1997, doi: 10.1518/001872097778543886.

[18] J. M. Logg, J. A. Minson, and D. A. Moore, “Algorithm appreciation: People prefer algorithmic to human judgment,” *Organizational Behavior and Human Decision Processes*, vol. 151, pp. 90–103, Mar. 2019, doi: 10.1016/j.obhdp.2018.12.005.

[19] B. J. Dietvorst, J. P. Simmons, and C. Massey, “Algorithm aversion: People erroneously avoid algorithms after seeing them err.,” *Journal of Experimental Psychology: General*, vol. 144, no. 1, pp. 114–126, 2015, doi: 10.1037/xge0000033.

[20] A. Kumar, T. Patel, A. S. Benjamin, and M. Steyvers, “Explaining Algorithm Aversion with Metacognitive Bandits,” *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43, no. 43, 2021.

[21] “GE jet engine bracket challenge” https://grabcad.com/challenges/ge-jet-engine-bracket-challenge

[22] E. Whalen, A. Beyene, and C. Mueller, “SimJEB: Simulated Jet Engine Bracket Dataset,” *arXiv:2105.03534 [cs]*, May 2021. Available: http://arxiv.org/abs/2105.03534

[23] J. Brookes, M. Warburton, M. Alghadier, M. Mon-Williams, and F. Mushtaq, “Studying human behavior with virtual reality: The Unity Experiment Framework,” *Behav Res*, vol. 52, no. 2, pp. 455–463, Apr. 2020, doi: 10.3758/s13428-019-01242-0.

[24] L. Chong, G. Zhang, K. Goucher-Lambert, K. Kotovsky, and J. Cagan, “Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice,” *Computers in Human Behavior*, vol. 127, p. 107018, Feb. 2022, doi: 10.1016/j.chb.2021.107018.

[25] R. F. Kizilcec, “How Much Information? Effects of Transparency on Trust in an Algorithmic Interface,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, 2016, pp. 2390–2395. Available: https://doi.org/10.1145/2858036.2858402

[26] C. J. Cai, J. Jongejan, and J. Holbrook, “The effects of example-based explanations in a machine learning interface,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Marina del Ray California, Mar. 2019, pp. 258–262. doi: 10.1145/3301275.3302289.

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

[28] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz, “Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff,” *AAAI*, vol. 33, pp. 2429–2437, Jul. 2019, doi: 10.1609/aaai.v33i01.33012429.