Interaction Design With Multi-Objective Bayesian Optimization

Yi-Chi Liao, *Aalto University, 02150, Espoo, Finland*

John J. Dudley and George B. Mo, *University of Cambridge, Cambridge, CB2 1PZ, U.K.*

Chun-Lien Cheng and Liwei Chan, *National Yang Ming Chiao Tung University, Hsinchu, 30010, Taiwan*

Antti Oulasvirta, *Aalto University, 02150, Espoo, Finland*

Per Ola Kristensson, *University of Cambridge, Cambridge, CB2 1PZ, U.K.*

Interaction design typically involves challenging decision making that requires designers to consider multiple parameters and careful tradeoffs between various objectives. This article examines how AI can facilitate the process of interaction design by offloading some of the complex decision making required of designers. We study how multi-objective Bayesian optimization can be used to support designers when creating a tactile display for smart watches. We present the results of a study that explores how such human–AI collaboration afforded by multi-objective Bayesian optimization can be exploited by designers, and the advantages and disadvantages this solution offers over conventional design practice.

Interaction design is challenging and a part of this challenge is the complexity of the design space, which is only exacerbated in pervasive computing applications. The user's experience and performance when interacting with a system is often governed by a large number of configurable design parameters. Adding still further complication is the fact that design objectives, such as performance, accuracy, or comfort, may be in tension or in direct conflict with each other and thus demand explicit or implicit tradeoffs to be decided by the designer. However, exhaustively examining the design space and assessing the impact of various design configurations is rarely feasible and in practice designers rely on their past experiences, design know how and established conventions to arrive at a particular design.

For example, consider the task of designing a distinguishable set of vibration-based notifications on a smart watch. Intuitively, a distinguishable set of vibration cues could be delivered by simply selecting distinct combinations of vibration durations and amplitudes. If we assume that the objectives in this design problem are to maximize cue recognition accuracy as well as the total number of distinguishable cues, how exactly should one go about methodically exploring the space of possible designs?

One conventional *ad hoc* approach to design space exploration is to manually select promising design candidates and sequentially evaluate these with participants. However, this process is difficult to perform systematically, in particular when the design space is large and may contain multiple competing objectives. Further, there is a risk that the design space exploration process inadvertently absorbs the biases and subjective preferences of the designer.

This article explores an alternative approach where the designer is partnered with an AI agent that intelligently proposes designs for evaluation. We study designers' experiences when interacting with such an AI agent to design a pervasive computing user interface—a tactile display for a smart watch. We focus on the designer's experience, as opposed to the end-user's experience, as we see a critical need to preserve a designer's appreciation of the design space. This focus reflects the fact that novel interactions may be designed in isolation but must typically be integrated into an actual application. At this integration stage it is useful if the designer can exploit their appreciation of the design space to understand how particular design decisions might be influenced by the much broader demands of the application.

1536-1268 © 2023 IEEE
Digital Object Identifier 10.1109/MPRV.2022.3230597
Date of publication 23 January 2023; date of current version 27 March 2023.

Among a potential range of techniques that may be suitable for AI-assisted interaction design we investigate multi-objective Bayesian optimization (MOBO). Bayesian optimization is a method for performing optimization on black-box functions. In interaction design we can view the mapping between the design parameters and the quality or performance of the design as such a black-box function. Bayesian optimization constructs a surrogate model of this unknown function and leverages this model to intelligently determine a promising new point in the design space to evaluate. Each new observation of the design space serves to improve the surrogate model. As such, Bayesian optimization might be particularly suitable for guiding interaction design for three reasons. First, it efficiently pursues promising designs while ignoring demonstrably poor regions of the design space. Second, it relies on very few initial assumptions compared to other optimization methods. Third, it is able to accommodate the high levels of noise typically inherent in observations of human behavior.

However, despite its suitability, MOBO is not widely used in interaction design. To address this gap, and to highlight the potential of MOBO in interaction design, we study the experience and performance of designers working in collaboration with MOBO. Further, we go beyond simply exploring whether MOBO is useful by also investigating the ease or difficulty with which designers can conceptualize and integrate MOBO into a design problem. We examine the relative merits and deficiencies of MOBO-supported human–AI collaboration by asking designers to complete the same design task using both their own preferred design approach as well as with the assistance from MOBO. This approach enables the study participants to directly reflect on the experience of using MOBO compared to the design process that they might otherwise use. The design problem we pose to designers is to design a maximally expressive and distinguishable set of vibration cues for delivering smart watch notifications. A successful design will rely on the ability of the designer to carefully balance competing objectives and a relatively large design space.

Three key takeaways emerge from the results of this study. First, a MOBO procedure results in designs that exhibit very similar performance as designs generated by designers using their own self-elected design strategies. Second, MOBO significantly reduces designers' perceived overall workload while successfully assisting the designers in identifying promising design candidates. Third, designers may become detached from the design process when typical aspects of their role are subsumed by MOBO.

Overall, MOBO appears to be a promising complementary AI-assisted design method suitable when design problems are complex and have multiple competing objectives. However, HCI research should study methods that help designers retain ownership and agency in the process.

BACKGROUND

Bayesian optimization is a machine learning technique for performing optimization of black-box functions. It works by constructing a surrogate model of the unknown function and selecting new points to evaluate with reference to this model. Each new observation improves the estimate of the model and thereby the understanding of where promising points are likely to lie. Bayesian optimization has been successfully applied to HCI design problems to refine interfaces,[1] and to support image[2,3] and animation customization.[4] Our work sits within a much broader body of research focused on AI-supported decision making.[5,6,7]

These prior studies used Bayesian optimization for a single objective. However, it is also possible to perform MOBO. As is true for all forms of multi-objective optimization, the outcome is no longer a single optimum but rather a set of optima representing various tradeoffs between the different objectives. This set of optima is referred to as the Pareto front. In the context of interaction design, the Pareto front represents a set of possible designs for which one objective cannot be improved without degradation of another objective. For example, a design involving a large set of distinct vibration-based notifications will maximize the potential for transferring information but may also give rise to more frequent misrecognitions. Conversely, a design with fewer distinct cues may support a high recognition accuracy but may limit the amount of information that can be transferred. Both designs may sit on the Pareto front and reflect optimal operating points exposing different tradeoffs in the design objectives.

The general form of the MOBO procedure is summarized in Figure 1. The designer initializes the procedure by determining a suitable parameterization for the design problem as well as the relevant design objectives. When the user-in-the-loop optimization process begins, new designs are presented and evaluated by the user. The results of these evaluations are fed back to refine the surrogate model and improve the selection of new designs. After some fixed number of iterations or set time period, the designer can inspect the surrogate model and extract the designs corresponding to the Pareto front.

The study presented in this article applies MOBO to the task of designing a wearable haptic display.

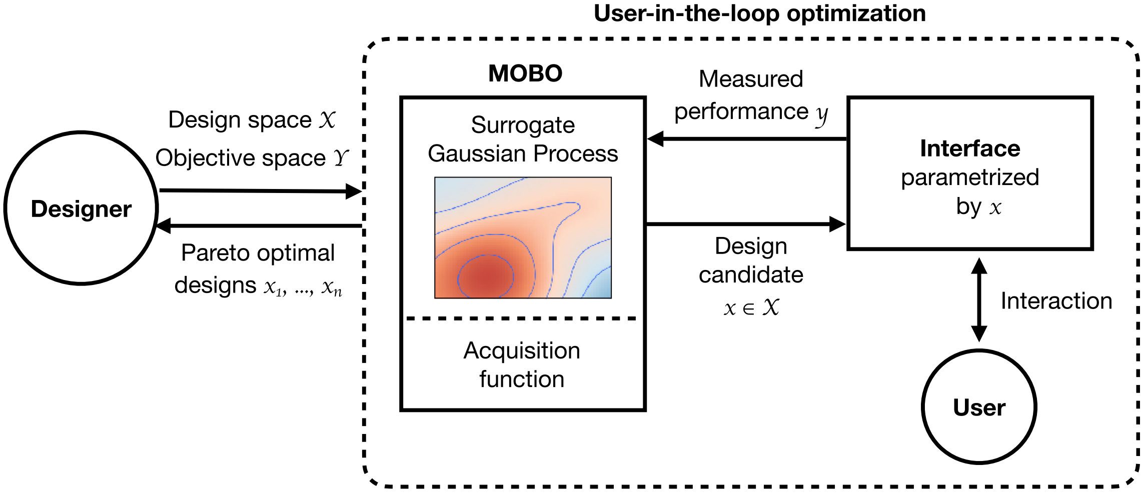

User-in-the-loop optimization
Designer
Design space X
Objective space Y
Pareto optimal designs $x_1, ..., x_n$
MOBO
Surrogate Gaussian Process
Acquisition function
Measured performance y
Design candidate $x \in X$
Interface parametrized by x
Interaction
User

FIGURE 1. Illustration of the MOBO procedure. The designer initializes the process by defining the design parameters (X) and the objectives (Y), and then commences the user-in-the-loop optimization. The MOBO procedure will propose design candidates, $x \in X$, for evaluation in the search for the Pareto-optimal parameter settings. The interaction technique behaviors are updated for each x and the user's performance and/or subjective experience is measured and translated into objective values y. The MOBO procedure updates its proxy model—a Gaussian process model based on the observed $\{x, y\}$ sets—which it then uses to propose a new design candidate. Once the optimization is complete, the designer gathers and selects from among the Pareto-optimal designs.

Investigating and optimizing the information transmission via haptic sensation has been a long-standing goal in haptic research. The problem has gained increasing attention with the emergence of smart watches.[8,9,10,11] Prior work has investigated a single vibrotactor generating vibrations with various durations, frequencies, and amplitudes[12,13,14] and delivering temporal-spatial patterns on skin with 2D tactors.[10,15,16]

Recent work by Chan et al.[17] conducted a related user study in which 40 novice designers were asked to create optimal designs for a 3D touch interaction either manually or by an optimizer-led approach using MOBO. Although the research protocols between this workshop study and Chan et al.'s work[17] is similar, there are several major differences in the experimental setup and the study goals. Chan et al.[17] investigated the benefits of MOBO for novice designers, where they searched for Pareto-optimal designs by assessing the performance of the interaction designs generated on themselves without external study participants. However, our study examines MOBO when applied in a scenario closer to what a typical designer would do in terms of evaluating an interaction design with several study participants. We only invited experienced designers and HCI researchers to take on the role of the designer, and we provided study participants to the designers during the entire design process for the designer to evaluate the designs generated.

STUDY

The study sought to answer two key questions. First, to what extent can MOBO be applied by designers to support their design work? Second, how does a MOBO procedure compare to the standard practice of designers? The study, which was conducted as a workshop, engaged participants in the task of designing the vibration cues for a haptic wearable display. The target participant group for the workshop was individuals with some experience of interaction design. For clarity, we subsequently refer to these participants as *designers*. Each designer was allocated two further participants who served as a proxy for users that the designer could use to test designs with.

The workshop was structured such that each designer completed the same design exercise twice: first using their own preferred design strategy, and second in collaboration with MOBO. Given the workshop approach, we focus primarily on observations regarding the designers' experience of the design procedure as opposed to the quality of the design outcomes.

Design Task

Designers were asked to tackle a classic problem in human–computer interaction—designing vibration cues for a haptic wearable display. One possibility for conveying different messages to the user with a single tactor is designing a set of vibrations that contain unique combinations of vibration duration and amplitude. More unique combinations allows for more messages to be conveyed. However, at some point the different messages become difficult to distinguish. As messages become more difficult to distinguish, more recognition errors will occur.

We purposefully constrained this design space by restricting the design problem to selecting an appropriate range for vibration duration and amplitude, as well as the number of distinct levels of duration and amplitude over that range. We fixed the maximum duration of vibration to 1 second and the maximum amplitude to 1.45 g. We then parameterized the design of the wearable tactile display according to the four design parameters (X) (and permitted ranges/values) summarized in the following.

- Minimum duration time of the vibration [50 ms, 950 ms].
- Number of discrete vibration duration levels $\{1, 2, 3, 4\}$.
- Minimum amplitude of vibration [0 g, 1.45 g].
- Number of discrete amplitude levels $\{1, 2, 3, 4\}$.

Designers participating in the workshop were presented with the following design brief. "You are asked to design the vibration cues for a newly released smart watch. Your task is to use the four design parameters available to create a set of vibrations that can achieve both high recognition accuracy and high information transfer rate. Other products on the market deliver up to three different vibration cues, and your final design is expected to at least outperform these competitors."

This brief refers to the two objectives (Y), which are to govern the optimization process: the information transfer (IT) rate and recognition accuracy. IT represents an estimate of the channel capacity for a specific stimuli set, that is, the effective bits of information transferred per stimuli (see Tan et al.'s work[18] [Section II, A] for the calculation of IT). We calculate recognition accuracy as $\frac{n_{\text{correct}}}{n}$ where n_{correct} is the number of trials in which the correct response is matched with the given stimulus, and n is the total number of trials.

Participants

Eight designers were recruited for the study (age 23–31; five females). Four designers were Ph.D. students recruited from local universities, and all of them conducted research in human–computer interaction. One designer worked as a professional user experience designer and had rich experience in conducting user research and data analysis. The remaining three designers were master's students majoring in design programs at local universities and they all had prior experience in running user studies.

Sixteen proxy users (age 21–33; seven females) were also recruited for the workshop so that each designer was assigned two dedicated study participants. Both the designers and the study participants were compensated based on the number of hours they participated in the workshop. The hourly compensation was € 11.

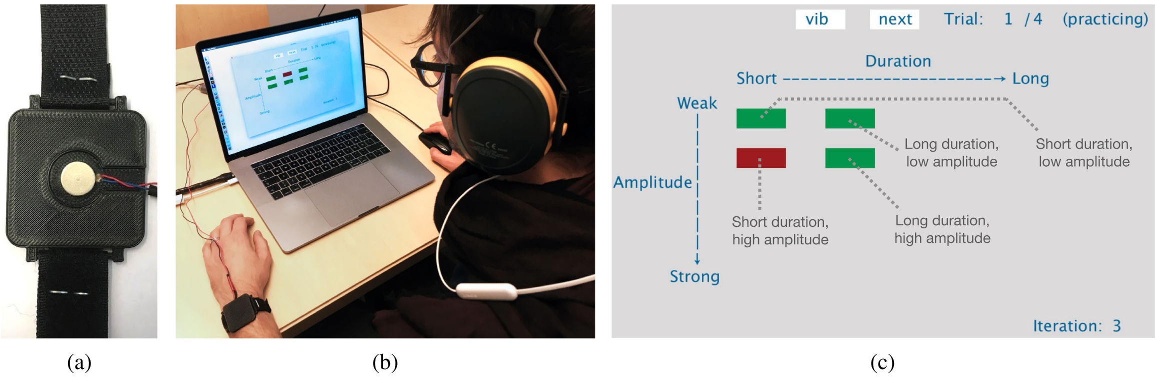
Procedure

The designers completed the same design exercise using both the MOBO procedure and their own chosen design procedure. For clarity, we refer to these two alternative design procedures as distinct conditions, even though in practice the designer-elected workflow may have been different for each designer. Four of the designers completed the workshop using the MOBO procedure first and their chosen design procedure second, while the remaining designers undertook the conditions in reverse order. This ensured that the conditions were counterbalanced in an attempt to control for learning effects.

The designers were each given a prototype of the wearable haptic smart watch [see Figure 2(a)] and assigned two dedicated additional study participants who served as proxy end users to test their designs with. The study setup was as pictured in Figure 2(b). The user interface shown in Figure 2(c) allowed designers to both instruct their two study participants about the mapping between the vibration cues and the notification intent, as well as to capture the ability of the two study participants to recognize cues. The set of possible cues for the current designs was displayed as a grid of boxes. A box further toward the right-hand side represented a cue with a longer vibration duration while a box further toward the bottom represented a cue with a larger vibration amplitude. The interface supported both a practice and a test mode. In the practice mode, participants were randomly presented with a specific cue and the corresponding box in the interface would turn red. In the test mode, participants clicked on the box that they believed corresponded to the cue presented.

The designers were given three hours for each condition. If needed, they could ask the workshop facilitator (one of the authors) for technical support and clarification about the design brief. The procedure differed slightly for each condition as detailed in the following.

Designer-elected procedure: The designers could directly choose a particular set of parameter values in


vib
next
Trial: 1 / 4 (practicing)
Duration
Short
Long
Weak
Amplitude
Strong
Long duration, low amplitude
Short duration, low amplitude
Short duration, high amplitude
Long duration, high amplitude
Iteration: 3

(a) (b) (c)

FIGURE 2. (a) Haptic display prototype. (b) Study setup with participant wearing the prototype and interacting with the user interface. (c) Detail of the user interface.

the interface and present these to their two study participants. Each design iteration would include both practice and test modes. When the test session of a particular design configuration was completed, the designer was shown the recognition accuracy and the number of cues. The designers familiarized themselves with the task and devised a study plan in the first hour of the session. The study plan was then executed in the remaining two hours. At completion, the designers were asked to specify their preferred final design.

MOBO procedure: The MOBO procedure illustrated in Figure 1 was implemented as a set of API calls that could be used by designers. The MOBO procedure itself employed the correlated expected improvement in Pareto hypervolume acquisition function proposed by Shah and Ghahramani.[19] Hyperparameters were tuned at each step of the optimization process by maximizing the log likelihood. Designers configured the procedure by specifying the parameterization, objectives, and some basic hyperparameters, and then initiated the user-in-the-loop optimization process. We provided a simple method that could combine the observations obtained from the two dedicated participants into a single model for extraction of the Pareto-optimal designs. The workshop facilitator (one of the authors) spent an hour introducing designers to the MOBO procedure and guiding them through the setup of the optimization process. The designers then had two hours to complete the design task, at the end of which they selected one final design from the Pareto frontier. The final Pareto frontier was generated from all collected data and presented as a 2D plot with each axis corresponding to one of the objectives.

After completing both conditions the designers were presented with their derived designs and their performances. The designers then completed a NASA-TLX and a system usability scale (SUS) questionnaire. Finally, the designers participated in an interview in which they were invited to reflect on the advantages and disadvantages in the two conditions' different procedures. In total, each designer was engaged for approximately seven hours.

RESULTS

Overall the designs arrived at using either a designer-elected procedure or the MOBO procedure was very similar. The mean accuracy and number of distinct vibration cues for the designs produced in the designer-elected workflow were 0.86 ($\mathrm{SD} = 0.08$) and 6.13 ($\mathrm{SD} = 1.36$), respectively. The mean accuracy and number of distinct vibration cues for the designs produced in the MOBO procedure were 0.88 ($\mathrm{SD} = 0.08$) and 6.13 ($\mathrm{SD} = 2.03$), respectively.

A major point of contrast between the MOBO procedure and designer-elected procedure is that the Pareto frontier generated by MOBO enables a structured interpretation of the influence of the design parameters. We observed that the optimal designs identified by MOBO had minimum vibration durations and minimum vibration amplitudes at the lower end of the feasible range. This aligns with intuition given that lower minimum vibration durations and amplitudes will produce more distinct individual cues. We also observed that the precise balance between accuracy and information transfer was chiefly influenced by the combined variation of the number of vibration duration levels and the number of vibration amplitude levels: reducing the number of levels for both parameters produced more accurate designs.

The designers used a variety of design strategies when choosing their own design procedure, which highlights the complexity of tackling the design problem using a conventional design procedure. In the following, we summarize the various design strategies that were applied by designers in the non-MOBO condition.

Divide-and-conquer and subsequently increasing the complexity of the task: Two designers (*D1* and *D2*) applied a divide-and-conquer approach by focusing on certain design parameters in isolation first and subsequently increasing the vibration cue count. They involved their two study participants throughout this process. Both designers determined the acceptable minimum duration and amplitude through testing with the two study participants as the first step. Then they tested 1 (duration level) × 2 (amplitude level) and 2 (duration level) × 1 (amplitude level) designs with the participants, which yielded a near perfect recognition rate but a relatively low cue count. The designers then gradually increased the number of vibration cues in the design until both the recognition rate and the cue count met the requirements.

Divide-and-conquer and subsequently decreasing the complexity of the task: D3 and *D6* applied a divide-and-conquer approach as described previously, but then subsequently gradually decreased the vibration cue count. First, they derived the minimum duration and amplitude by testing with their two study participants. Then, they tested the largest possible vibration set, 4 (duration level) × 4 (amplitude levels), yielding a high cue count but a low recognition rate. The designers then incrementally decreased the number of vibration cues and tested each design generated with the participants for every change. The design process stopped when a satisfactory recognition rate was achieved.

Divide-and-conquer and local search: D5 and *D7* spent 30 minutes deriving a starting design with a medium number of vibration cues: one designer started with a 3 (duration level) × 2 (amplitude level) design and the other designer started with a 3 (duration level) × 3 (amplitude level) design. These initial designs were relatively close to their final designs in terms of their design parameters. The designers followed a strategy similar to performing a "local search" where they fine tuned the design parameters until they were satisfied.

Self-evaluating approach: D8 largely evaluated the generated designs without involving their two study participants. After an hour of self testing, *D8* derived three final design candidates. *D8* then invited the two study participants to evaluate these final design candidates and thereafter selected the design candidate with the highest preference.

Focus group: D4 adopted a "focus group" approach. The designer allocated both the two study participants and themselves five minutes to create their own designs independently. Then, all three people (the two study participants and the designer) evaluated all the designs made by the others. Then the group discussed how to improve the design, followed by another round of evaluation using the same approach. After two iterations, the group narrowed down the selection to two final designs.

The total number of designs evaluated varied across designers (*D1:* 6; *D2:* 6; *D3:* 5; *D4:* 8; *D5:* 9; *D6:* 6; *D7:* 5; *D8:* 3).

Usability and Workload

We assess significant differences in the overall and subscale ratings for usability and perceived workload using a Wilcoxon signed-rank test based on participant matched samples. Figure 3 summarizes the results of the SUS questionnaire. The mean SUS score for the designer-elected procedure was 54.38 ($\mathrm{SD} = 15.51$) and 64.38 ($\mathrm{SD} = 13.48$) for the MOBO procedure. A Wilcoxon signed-rank test revealed no statistical difference between the overall SUS scores ($Z = -1.02, p > 0.05$).

Although the overall scores were not significantly different, there were statistical differences when examining individual questions. Based on a Wilcoxon signed-rank test, there were statistically significant differences in the responses to Q1 ($Z = -2.06, p < 0.05$) and Q4 ($Z = -2.97, p < 0.05$). This suggests that designers would indeed like to use the MOBO procedure (Q1) but that they would require more technical support (Q4).

Figure 4 summarizes the results of the NASA-TLX questionnaire. The mean workload for the designer-elected procedure was 62.67 ($\mathrm{SD} = 16.36$) and 45.17 ($\mathrm{SD} = 12.4$) for the MOBO procedure. A Wilcoxon signed-rank test revealed a significant difference between the overall workloads for the two conditions ($Z = -2.38$, $p < 0.05$). In other words, the MOBO procedure significantly reduced mental workload. Further, based on Wilcoxon signed-rank tests, the designer-elected strategy elicited statistically higher mental demand ($Z = -2.2$, $p < 0.05$), physical demand ($Z = -2.06, p < 0.05$), temporal demand ($Z = -2.37, p < 0.05$), and frustration ($Z = -4.53, p < 0.05$).

The SUS and NASA-TLX results show that the MOBO procedure delivered a usable alternative design process. Further, the MOBO procedure generally induced a lower cognitive load, frustration, and mental and physical demand than the designer-elected procedures.

User Experience

All designers agreed that the MOBO-assisted design procedure largely reduced the effort involved in interpreting the data and making decisions throughout the design process. As noted by *D1:* "The design space is very large. The manual design process would take a lot of time and

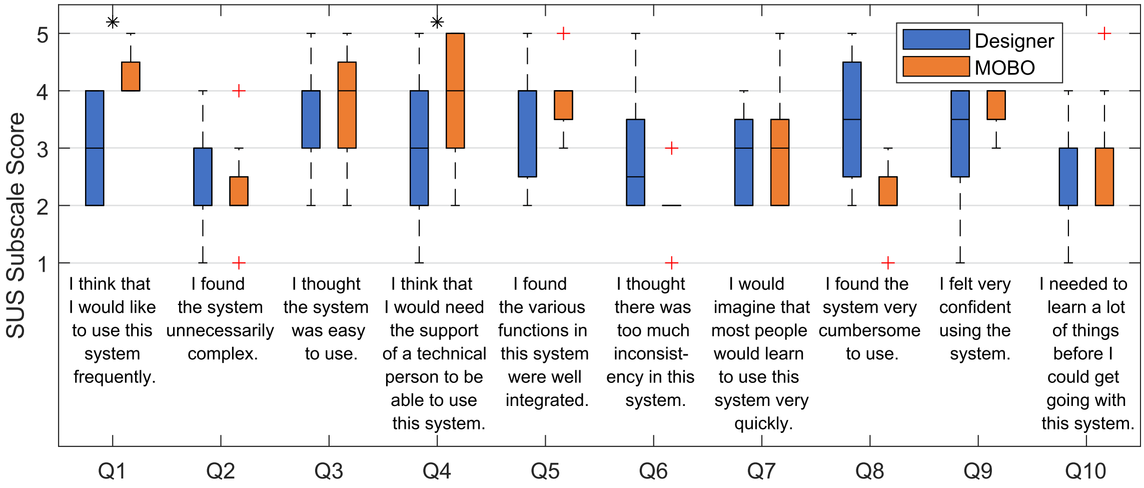


FIGURE 3. Boxplots showing the ratings from the eight designers for the SUS questionnaire, comparing the design-elected workflow with the MOBO procedure. The red crosses mark outliers, which are defined as beyond $Q_{1/3} \pm 1.5 \times (Q_3 - Q_1)$. The star ($*$) symbol indicates significant difference at $p < 0.05$. Significant differences were observed in Q1 and Q4 suggesting that designers would be more willing to use the MOBO procedure frequently than their own selected design strategy, but that they would be more likely to require technical support.

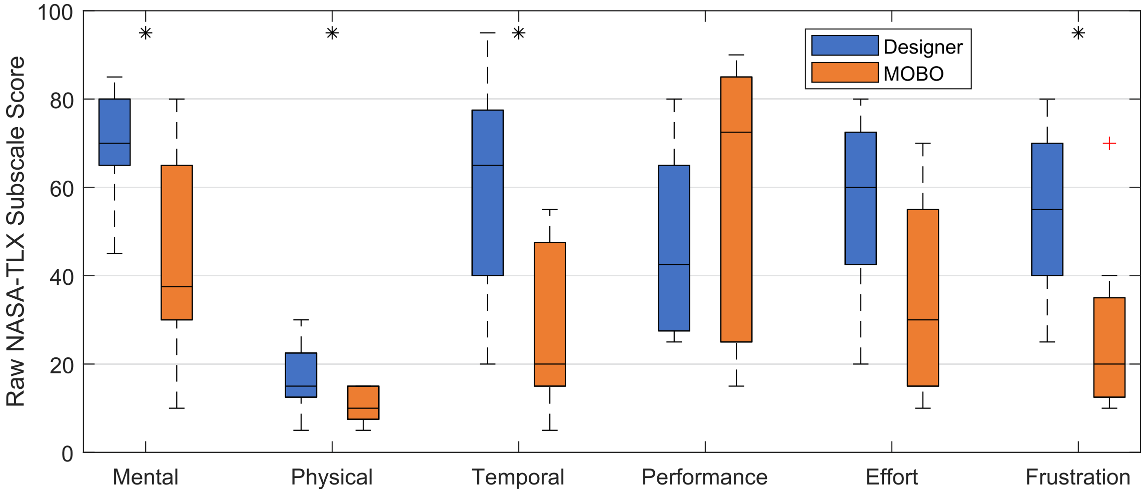


FIGURE 4. Boxplots showing the ratings from the eight designers for the NASA-TLX questionnaire, comparing the manual designer-elected workflow with the MOBO procedure. The red crosses mark outliers, which are defined as beyond $Q_{1/3} \pm 1.5 \times (Q_3 - Q_1)$. The one-star ($*$) symbol indicates significant difference at $p < 0.05$. The MOBO procedure yielded significantly lower mental, physical, and temporal demand, and significantly lower frustration.

effort to explore until reaching an acceptable design. [...] [MOBO] removed that effort of making decisions and trial-and-errors." *D8* shared a similar perspective: "In the manual design process, I needed to carefully consider tuning the design so that there will be an improvement. It is a demanding process and I constantly felt uncertain. However, MOBO just did that for me and I'm happy with the final results." *D2* also pointed out that the MOBO procedure helped to reduce not only the mental load but also the physical load: "[With MOBO] I did not need to manipulate the interface and the device, nor interact with the participants much. I simply needed to instruct the participants what to do, and the results would be generated, which is a big advantage."

The designers agreed with the benefits of having a series of proposed designs, as represented by the

Pareto front. *D1* observed: "If I changed my weights of the objectives and wanted to search for another design, I might need to invest another 30 minutes to reach that point. [The output of the MOBO procedure] showed all the designs along the line (Pareto front) and I could just pick one from them. From this perspective, I find [the MOBO procedure] much more efficient because it searches not just one final outcome but multiple." *D5* further mentioned: "I set some kind of priority at the beginning of the design. For example, the recognition rate is more important than the information transfer, and I want to achieve 95% of accuracy. However, during the [designer-elected] design process, I might gain new knowledge about the interaction, and would like to change the weight of the two objectives, which would force me to change the direction of search. The [MOBO procedure] can avoid this kind of hassle because it explores all the directions and provides all the possibilities."

All designers pointed out that their self-selected approaches induced higher frustration and temporal demand. *D4* stated: "This search can go on forever. I can always change something and lead to a different performance. I always feel uncertain, not knowing if this change will improve or not, and this is frustrating. Also, because I need to deliver a design within the certain amount of time, so I was somehow stressed." *D1* also shared that: "I was not sure if this design is good enough, so I felt it is more temporal demanding. On the other hand, when using [MOBO], I simply needed to assign one hour [...] to each participant and collected the results. It is much simpler and relaxing."

Overall, all the designers provided positive feedback to the final designs derived by the MOBO procedure. *D5* said, "I am surprised to see the results [produced by MOBO] are so good. It is well aligned with my own design. Also, the whole Pareto front looks promising to me and in line with my expectation." Further, *D7* used the Pareto-optimal designs generated by the MOBO procedure as a baseline to compare to: "Checking the designs made by [the MOBO procedure] is more like a reassuring step. To be honest, I trust the results made by the system more than the ones made by me. The Pareto front indicated a very systematical search."

In addition to the positive aspects of the MOBO procedure, the designers also highlighted several drawbacks. First, setting up the MOBO procedure required some level of programming experience. However, this issue can be mitigated if the designer is supported by a developer or potentially resolved by developing more elaborate tools for MOBO-assisted design in the future. As *D3* pointed out: "As a developer, I do not find major difficulty of using [MOBO], but I would assume a designer without coding background will need some kind of technical support." *D8* expressed a very similar view. Another major identified drawback of adopting the MOBO procedure is losing the opportunity to receive qualitative feedback from users regarding a particular design. *D2* mentioned: "I might want to learn the feedback from the participants, such as how do they feel about this design, and how can I improve from that. But [the MOBO procedure] did not give me this possibility." *D5* also mentioned: "I felt I lost the involvement if I fully rely on the [MOBO-assisted procedure]. I fully trusted the final results; they look promising and reasonable to me. Still, I would appreciate to talk to the participants and learn from them how they felt."

DISCUSSION AND CONCLUSION

Overall, the final designs produced by the MOBO procedure were found to be comparable in terms of performance to those generated by a designer-elected procedure. However, we found that the MOBO procedure significantly reduced the designers' perceived workload, which was also echoed by the qualitative data we gathered from the interviews.

The variety of approaches taken by the designers when they were allowed to choose their own design strategy demonstrates that there is no single obvious design strategy that can generate designs that guarantee any performance specification. We conjecture, as a consequence, the designers had to spend time and effort in devising a specific strategy for the design problem. In contrast, the MOBO procedure provided a single systematic approach for tackling the design problem. The designer was thus freed from the burden of conceiving a strategy and selecting a specific study plan, as the optimizer lead the generation of new designs to explore. Although the results of this study are broadly in line with Chan et al.'s work,[17] the investigation presented in this article of MOBO versus the designer's own selected design strategy highlights the additional cognitive burden the designer encounters when devising a custom experimental methodology. Further, we more precisely examine the experience of the designer as opposed to the merged designer/end user that was the subject of Chan et al.'s study.

There were primarily two downsides to using the MOBO procedure. First, it requires some experience in programming or the ability to call on a developer to support the setup of the system. This added complexity may in itself have negatively impacted the experience of using the MOBO procedure. Second, fully relying on a MOBO procedure may lead to the designer becoming detached from study participants and unable to utilize subjective feedback to drive the design process. Our work therefore provides further motivation for the research community to develop human–AI interactions that promote positive synergy.[20]

The MOBO procedure in this article is an example of how AI can be fruitfully used as a partner with a designer to exploit a complex design space with competing objectives for a pervasive computing application. We view it as highly encouraging that the AI-assisted designs closely match the outcomes of the designers' plethora of self-elected approaches in the study, which is a strong indication that partnering a designer with an AI can at the very least result in comparable results and with a significantly reduced perceived workload. However, we also note the many avenues for future work.

First, we see fruitful future work in *tool design* that alleviates the need for programming expertise. Further, such tools should ideally incorporate techniques for clearly explaining designs proposed to users, their tradeoffs and implications, and the inherent uncertainty associated with their measured performance.

Second, there is a challenge in avoiding the MOBO procedure resulting in the designer becoming too detached from study participants. This well-known problem in automation is expected but will need to be tackled for a MOBO procedure to ultimately achieve widespread adoption.

Third, pervasive computing user interface designs, in general, are particularly challenging as they often rely on context and/or uncertain sensing. It would be interesting to consider a more complex design problem, for example, an interface that, in part, relies on working within a specific environment for successful interaction.

ACKNOWLEDGMENTS

The authors would like to thank all the participants for attending the user study. The work of authors Yi-Chi Liao and Antti Oulasvirta were supported by the Academy of Finland projects BAD and Human Automata. The work of author Antti Oulasvirta was supported by the Finnish Center for AI, an Academy of Finland flagship. The work of authors John J. Dudley and Per Ola Kristensson were supported by a Grant from the Engineering and Physical Sciences Research Council under Grant EPSRC EP/S027432/1. The work of author Liwei Chan was supported by the Ministry of Science and Technology of Taiwan under Grant MOST109-2628-E-009-010-MY3. This work was supported by the Department of Information and Communications Engineering (Aalto University). This work involved human subjects or animals in its research. The study was carried out according to the ethical principles of research with human participants and ethical review in human sciences in Finland governed by the Finnish National Board on Research Integrity.

REFERENCES

1. J. J. Dudley, J. T. Jacques, and P. O. Kristensson, "Crowdsourcing interface feature design with Bayesian optimization," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–12, doi: 10.1145/3290605.3300482.
2. Y. Koyama, I. Sato, D. Sakamoto, and T. Igarashi, "Sequential line search for efficient visual design optimization by crowds," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 48:1–48:11, Jul. 2017, doi: 10.1145/3072959.3073598.
3. Y. Koyama, I. Sato, and M. Goto, "Sequential gallery for interactive visual design optimization," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 88:88:1–88:88:12, Jul. 2020, doi: 10.1145/3386569.3392444.
4. E. Brochu, T. Brochu, and N. de Freitas, "A Bayesian interactive optimization approach to procedural animation design," in *Proc. ACM SIGGRAPH/Eurograph. Symp. Comput. Animation, Goslar, DEU: Eurograph. Assoc.*, 2010, pp. 103–112.
5. V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, "Towards a science of human-AI decision making: A survey of empirical studies," 2021, *arXiv: 2112.11471.*
6. H. Liu, V. Lai, and C. Tan, "Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making," *ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, Oct. 2021, Art. no. 408, doi: 10.1145/3479552.
7. 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 *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 295–305, doi: 10.1145/3351095.3372852.
8. R. H. Gault, "Progress in experiments on tactual interpretation of oral speech," *J. Abnorm. Psychol. Social Psychol.*, vol. 19, no. 2, pp. 155–159, 1924.
9. J. H. Kirman, "Tactile perception of computer-derived formant patterns from voiced speech," *J. Acoustical Soc. Amer.*, vol. 55, no. 1, pp. 163–169, 1974, doi: 10.1121/1.1928145.
10. S. C. Lee and T. Starner, "Buzzwear: Alert perception in wearable tactile displays on the wrist," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2010, pp. 433–442, doi: 10.1145/1753326.1753392.
11. M. Matscheko, A. Ferscha, A. Riener, and M. Lehner, "Tactor placement in wrist worn wearables," in *Proc. Int. Symp. Wearable Comput.*, 2010, pp. 1–8.

12. H. Z. Tan, N. I. Durlach, W. M. Rabinowitz, C. M. Reed, and J. R. Santos, “Reception of morse code through motional, vibrotactile, and auditory stimulation,” in *Perception and Psychophysics*. Berlin, Germany: Springer, 1997, pp. 1004–1017.
13. D. Ternes and K. E. Maclean, “Designing large sets of haptic icons with rhythm,” in *Proc. 6th Int. Conf. Haptics: Percep., Devices Scenarios, EuroHaptics*, 2008, pp. 199–208, doi: 10.1007/978-3-540-69057-3_24.
14. S. Brewster and L. M. Brown, “Tactons: Structured tactile messages for non-visual information display,” in *Proc. 5th Conf. Australas. User Interface*, 2004, vol. 28, pp. 15–23.
15. J. Lee, J. Han, and G. Lee, “Investigating the information transfer efficiency of a 3x3 watch-back tactile display,” in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, 2015, pp. 1229–1232, doi: 10.1145/2702123.2702530.
16. Y.-C. Liao, Y.-L. Chen, J.-Y. Lo, R.-H. Liang, L. Chan, and B.-Y. Chen, “EdgeVib: Effective alphanumeric character output using a wrist-worn tactile display,” in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, 2016, pp. 595–601, doi: 10.1145/2984511.2984522.
17. L. Chan et al., “Investigating positive and negative qualities of human-in-the-loop optimization for designing interaction techniques,” in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2022, Art. no. 112, doi: 10.1145/3491102.3501850.
18. H. Z. Tan, S. Choi, F. W. Y. Lau, and F. Abnousi, “Methodology for maximizing information transmission of haptic devices: A survey,” *Proc. IEEE*, vol. 108, no. 6, pp. 945–965, Jun. 2020.
19. A. Shah and Z. Ghahramani, “Pareto frontier learning with expensive correlated objectives,” in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1919–1927. [Online]. Available: http://proceedings.mlr.press/v48/shahc16.html
20. A. Campero, M. Vaccaro, J. Song, H. Wen, A. Almaatouq, and T. W. Malone, “A test for evaluating performance in human-computer systems,” 2022, *arXiv:2206.12390*.

YI-CHI LIAO is a Ph.D. candidate in the User Interfaces group at Aalto University, 02150, Espoo, Finland. He has interned at Meta Reality Labs. His research focuses on using Bayesian optimization to assist UI design and reinforcement learning to model human interactions with physical interfaces. Liao received his master’s degree from National Taiwan University, Taipei, Taiwan. Contact him at yi-chi.liao@aalto.fi or visit his personal website at http://yichiliao.com for more information about his research.

JOHN J. DUDLEY is an associate teaching professor in the Department of Engineering at the University of Cambridge, Cambridge, CB2 1PZ, U.K., and a postdoctoral associate of Jesus College, University of Cambridge. His research focuses on the design of interactive systems that dynamically adapt to user needs and behaviors. He is the corresponding author of this article. Contact him at jjd50@cam.ac.uk.

GEORGE B. MO received his M.Eng. degree in engineering from the University of Cambridge, Cambridge, CB2 1PZ, U.K. He was a member of the Intelligent Interactive Systems group at the University of Cambridge during the execution of this work. Contact him at georgemo535@gmail.com.

CHUN-LIEN CHENG was a student at National Yang Ming Chiao Tung University, Hsinchu, 30010, Taiwan, during the execution of this work. Contact him at jimmy61209@gmail.com.

LIWEI CHAN is an associate professor in the Department of Computer Science at the National Yang Ming Chiao Tung University, Hsinchu, 30010, Taiwan. His research interests include human–computer interaction, interaction design for AR/VR, and haptic user interfaces. Contact him at liweichan@cs.nycu.edu.tw.

ANTTI OULASVIRTA is a professor of user interfaces with Aalto University, 02150, Espoo, Finland. He leads the interactive AI programme at the Finnish Center for AI. His research focuses on computational methods in human–computer interaction, including interactive ML, user modeling, and simulation. Contact him at antti.oulasvirta@aalto.fi.

PER OLA KRISTENSSON is professor of interactive systems engineering in the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K., and a fellow of Trinity College, Cambridge. He leads the Intelligent Interactive Systems group, which belongs to the Engineering Design Centre. He is also a co-founder and co-director of the Centre for Human-Inspired Artificial Intelligence, University of Cambridge. Contact him at pok21@cam.ac.uk.