The Impact of Expertise in the Loop for Exploring Machine Rationality

Changkun Ou
LMU Munich
Germany
research@changkun.de

Sven Mayer
LMU Munich
Germany
info@sven-mayer.com

Andreas Butz
LMU Munich
Germany
butz@ifi.lmu.de

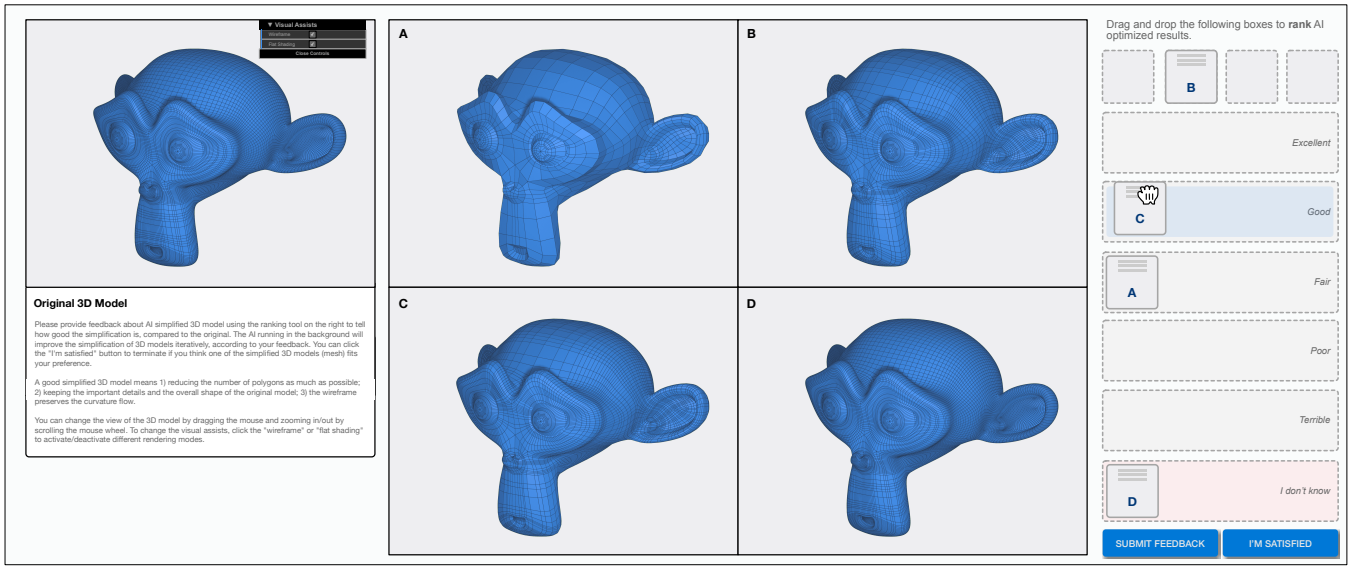


Figure 1: The interface for optimizing 3D model simplification using an expert in the loop. In each iteration, the interface presents four 3D models. Participants can drag and drop the top right blocks to a suitable rating region to provide a ranking at submission. Each of the regions can contain multiple blocks. Blocks can be put into "I don't know" to express an incomplete preference. Participants can indicate their satisfaction by terminating the optimization loop. To inspect the 3D model quality, they can zoom in/out, pan, and rotate the 3D models simultaneously using a mouse.

ABSTRACT

Human-in-the-loop optimization utilizes human expertise to guide machine optimizers iteratively and search for an optimal solution in a solution space. While prior empirical studies mainly investigated novices, we analyzed the impact of the levels of expertise on the outcome quality and corresponding subjective satisfaction. We conducted a study (N=60) in text, photo, and 3D mesh optimization contexts. We found that novices can achieve an expert level of quality performance, but participants with higher expertise led to more optimization iteration with more explicit preference while keeping satisfaction low. In contrast, novices were more easily satisfied and terminated faster. Therefore, we identified that experts seek more diverse outcomes while the machine reaches optimal results, and the observed behavior can be used as a performance indicator for human-in-the-loop system designers to improve underlying models. We inform future research to be cautious about the impact of user expertise when designing human-in-the-loop systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Interaction paradigms*; • **Computing methodologies** → *Active learning settings.*

KEYWORDS

human-in-the-loop machine learning; adaptive human-computer interaction; rationality

ACM Reference Format:
Changkun Ou, Sven Mayer, and Andreas Butz. 2023. The Impact of Expertise in the Loop for Exploring Machine Rationality. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3581641.3584040

IUI '23, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0106-1/23/03.
https://doi.org/10.1145/3581641.3584040

1 INTRODUCTION

Human-in-the-loop (HITL) [51] optimization uses human expertise to improve machine capabilities. It optimizes system parameters according to human evaluation feedback and supports humans to obtain better outcomes in a variety of areas, such as co-creation [50], personalized recommendation [32], and decision-making [4]. When designing a system that involves human feedback, a frequent design decision is not to query absolute human ratings but to ask about the preferred option among a set of design alternatives [40, 71]. If this is done repeatedly, the system iteratively models the user's preferences from the given feedback to infer the next optimal set of options. From a machine perspective, these options represent curiosity regarding what humans might prefer.

When humans face a design problem that requires adjusting various system parameters for the desired outcome, it is tedious to tweak them without prior expertise with the system's behavior. To improve the feedback loop efficiency, one can substitute this process of choosing different parameters with adaptive *exploration* and *exploitation* using human choices. Among many existing approaches, the Bayesian optimization (BO) technique is frequently used and preferred [11, 12, 61]. With BO applied, the user interface (UI) can, for instance, present multiple design alternatives, from which users then make a decision [71]. The system automatically predicts the next best estimations and presents them again based on past choices. The process not only removes the user workload of tweaking parameters but is also expected to propose desired outcomes eventually [48]. An underlying assumption is that the system outcome could improve when more human expertise is involved in this interaction loop.

However, the system might not always be effective in achieving user satisfaction. There are several known reasons for this, such as context [57], timing [28], trustworthiness [35], cognitive biases [12], and unstable and contradicting preferences [53]. Specifically, in co-creation, a user is not always satisfied with the results generated by the machine due to the lack of practical creativity in the system [50]; in a personalized recommendation, the machine may converge to some fixed recommended content and cannot bring fresh ideas for users [32]; even in the process of AI-assisted decision-making, users may not hold enough trust in the results provided by an algorithm [18].

Although there are strategies to mitigate these subjective imperfections on the human side, such as improving transparency [35], interpretability [19], and control [72], the reported dissatisfaction, lack of freshness, and trust remain subjective and are measured exclusively using subjective scales, Moreover, empirical studies also mainly report based on novice user groups [13, 40]. The impact of the involved expertise on the overall system outcome quality is rarely discussed. On the other hand, we can not easily assess the objective outcome quality reliably if the results partially depend on subjective concepts. Since the expertise involved plays an important role in the obtained human feedback, we investigate the following two research questions:

RQ1 *How do HITL optimization outcomes differ objectively when using preferential feedback from humans?*, and

RQ2 *What is the impact of the involved user expertise on the system outcomes and subjective satisfaction?*

In particular, we are interested to see *how the answers to these questions could provide insights for designing future HITL systems.* To cover a spectrum of different application domains, we consider text summarization [64], photo color enhancement [39–41], and a 3D model simplification task [21, 31, 53] to evaluate the relation between user expertise, satisfaction, and system outcome quality when interacting with an intelligent system. Figure 1 shows one of our ranking interfaces for HITL optimization. Our selected tasks are not only challenging to design algorithms for and to solve technically, as they require not only objective measurements but also subjective opinions [2, 45]. Therefore, we conducted a study to collect choice behaviors in a user group (N=60) with different levels of expertise in three contexts to assess the overall interaction and optimization loop. We also asked about their subjective satisfaction regarding the final system outcomes.

As a key result, we evaluate the connection between user expertise, subjective satisfaction, and the quality of the system outcomes in an interactive feedback loop. Our evaluation indicates that novice subjects can produce an equal outcome quality even faster (and be more satisfied with it) than those with higher expertise. The main contribution of this paper is an empirical investigation of the impact of involved human expertise on the overall HITL optimization performance. We also discuss design implications and potential future directions to consider the impact of user expertise in HITL optimization applications.

2 BACKGROUND AND RELATED WORK

To start, we briefly discuss the state-of-the-art approaches for modeling human feedback iteratively using Bayesian optimization. Then, we overview prior literature in the social sciences, mainly psychology, to resolve the potential ambiguity regarding the terms "satisficing" and "expertise" and approaches to quantify them. They serve as the foundation of our problem description and support our assessment of user expertise and satisfaction in HITL settings.

2.1 Modeling Preference from Human Feedback

Human-in-the-loop optimization outcomes depend on the machine algorithm capability as well as the preferential feedback expressed by a human user. Studies on the term "preference" appear in many disciplines. For clarity in the subsequent discussions, in this paper, we follow Hausman [27] in their counterargument against eliminating preference using choice [26] and acknowledge the existence of *preference*. In our use of terms, shown in Figure 2, preference is a subjective concept representing an impermeable and unobservable state of an individual mind. A preference may or may not be present when the individual encounters multiple given *options*. A *choice* denotes the objectively observable actions of the individual that selects at least one option among the given ones, and *decision* or *judgment* reveals a subjective realization process from a preference towards a choice. A choice may not reflect the underlying preference due to external influences.

In existing theories regarding preferences in psychology and economics, theoretical models tend to infer preference from comparisons [65] and rely on basic axioms [1] of this preference logic: *completeness* and *transitivity*. The completeness axiom assumes the existence of preference, which guarantees that individuals can

always express their preference by choosing among at least two options; transitivity means that we can infer that A is preferred over C if we prefer A over B and B over C.

Although these axioms are convenient for a rigorous discussion of the logic of preferences, they are still strong assumptions that may easily be violated. Behavioral literature shows that choices are partly dominated by the context [57], and the transitivity axiom is not applicable when implicitly involving other judgments that were not previously considered. For example, when a human prefers A over B and B over C involving only one objective, they may implicitly involve another, previously unconsidered objective when choosing between A and C as a pair. As a result, C may be preferred over A. Moreover, the completeness axiom may be violated when the human thinks that "I don't know" or "I don't care" among two subjectively indistinguishable options, thus causing a *random choice*.

Since BO learns a *posterior* from human feedback, it aims to search for a maximum of an unknown function by exploring and exploiting the solution space. Therefore, it can propose examples using an *acquisition function*, ask the human to provide a choice, and then infer the underlying preference iteratively. When dealing with choices from pairwise comparisons, preferential Bayesian optimization (PBO)[1] as a specialized category of BO has received increasing development in recent studies [24, 33, 40, 49, 62]. While BO learns based on absolute rating utility (rate and assign a score to an option), PBO learns from human choice in pairwise comparisons according to Thurstone's law of comparative judgment [65]. To avoid the mentioned violations of the transitivity axiom, the recent extensions [7, 40, 62] to PBO transited from using a binary pairwise comparison to using a reasonable amount of options. These extensions can largely prevent violation of the transitivity axiom and infer more information at a time because they either consider choosing a set of options as winners among all given options [7, 40]; or provide a ranking of all given options, where options may share the same level of rank [62]. Note that more ranking elements may also increase the uncertainty for users to make imperfect decisions [51] due to increased workload. Thus, one should carefully consider the presented number of elements.

Although PBO has used pairwise comparisons to mitigate various issues regarding unstable human judgments, these current approaches could also violate preference axioms due to cognitive limitations. Tversky and Kahneman [67] have widely presented how heuristic biases might influence the choice behavior. External causes can also produce a considerable amount of noise in choice [34]. To overcome the violation of preference axioms, PBO has also considered handling noisy inputs [46] and guarantees theoretical convergence when dealing with unstable choices. Despite all these developments in PBO, we still observe two challenges in practice: The first challenge is that a human might change their objective during the integrative optimization, even using pairwise comparisons, because PBO assumes a fixed implicit underlying choice function which it can learn. Although PBO can deal with noisy inputs, another challenge is that it requires much more iterations to let the optimization converge. This is usually very costly when involving a human, and we also need to design the UI carefully to mitigate these issues and reduce user errors.

Therefore, there are several major design considerations to obtain reliable preference from choice: 1) the underlying learning algorithm should effectively deal with feedback uncertainty and noisy input of an individual being, 2) the objective of the user task should be provided to avoid incomplete preference, and 3) The user interfaces should support users to express their incomplete preferences explicitly.

[1]We use PBO as a more general term to represent a category of methods that infer preference from choice, in contrast to the specific approach by González et al. [24].

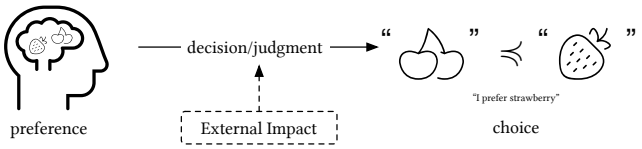


Figure 2: A decision process turns an internal preference state into an observable choice. The choice may not reflect the underlying preference due to external influences.

2.2 Bounded Rationality and Satisficing

Understanding the satisfaction of users when they are involved in a loop requires deeper insights from human psychological factors regarding bounded rationality and satisficing. Simon [63] first coined the term *bounded rationality* to describe the perceived information limits individual rationality. This observation provides a sufficient discussion base for interpretations regarding irrational decisions. As previously discussed, Tversky and Kahneman [67] emphasized one possible category of systematic errors from the cognitive perspective. In recent discussions [10, 23], researchers take a statistical perspective and underline that recurring noise could also contribute equally [34] to bounded rationality. This is met by matching behavior from the preference point of view, as bounded rationality appears in the decision or judgment process and causes the violation of the completeness axiom due to *satisficing*.

Satisficing is a "good enough" decision strategy [58, 63] that ends the search process when a certain threshold quality is met. When some of the presented options are subjectively acceptable, the effects of bounded rationality and satisficing cause the process to terminate with a sub-optimal outcome. An opposite decision tendency is called *maximizing*, where a final decision cannot be made without enough information. Schwartz et al. [60] provided evidence by assessing subjective happiness and individual differences in what people aspire to when they make decisions in various domains of their lives. People who use a maximizing strategy desire the best possible result. Although the authors did not find any strict causality for a maximizing strategy producing significantly lower satisfaction with life than satisficing, they argue that a maximizing decision strategy might constantly look for better objective outcomes. In modern recommender systems, for example, prior work [30] showed that a satisficing strategy leads to quicker selections and increased content viewing time. In contrast, subjects using a maximizing strategy spent significantly more time on selection activities. In comparison, subjects using satisficing decision strategies spent significantly less viewing time, regardless of subjective content quality.

The objective reasons why HITL optimization systems work differently for bounded rational human agents remain underexplored. Although previous psychology research correlated rationality with using satisficing and maximizing decision strategies, there is little discussion about what objective properties lead to the reported subjective dissatisfaction in this new context. Especially as the previously reported unsatisfactory results rarely evaluate the objective quality of optimized choice options while involving different levels of rationality, we also wonder if an unsatisfactory result has comparably lower quality or whether satisficing is sufficient to maximize a machine learner's capability. In addition, with proper selection on a task, the quality of a rational choice is also part of the consequence of human intelligence, known as "expertise". Concerning decisions using expertise, empirical research also reports that people with high expertise apply more criteria during their decision, especially clinic decisions [25], which proved less efficient and more correlated to a maximizing strategy. Still, it remains unclear what the satisfaction would be in this case.

2.3 Expertise in Context

To analyze the concept of expertise and quantify the impact of involved expertise, one of the most straightforward questions regarding expertise is: "what is an expert?" Depending on the domain context, there are different decompositions of the concept of expertise. In particular, Garrett et al. [22] describe six different dimensions regarding expertise, whereas Collins [15] suggests three dimensions and Kotzee and Smit [38] suggest only two dimensions based on social aspects. On a higher level, Bourne et al. [9] argues for interpreting expertise as a descriptive term that involves knowledge and skills, which are mental or cognitive concepts rather than physical talent. Therefore, tasks that might be physically quickly adapted and measured regarding efficiency are less suited to verifying the expertise involved.

To quantify the loosely defined concept "expertise," a range of theoretical models have been developed, e.g., by describing a game between a decision-maker and an expert who proposes options [42]. For our purpose here, we are interested in quantifying the level of expertise of a specific human within a particular context. Treem and Leonardi [66] propose to define 1) an observer who knows what it looks like and 2) an expert who has an objective communicative skill that outperforms the observer who can infer their expertise. Ooge and Verbert [52] further developed this concept and introduced a third metric for inferring expertise by using a preliminary task to measure a person's performance.

Because of the interpretation ambiguities and different arguments about proper measures in other contexts, instead of asking about an absolute level "is A an expert?", identifying a person with a comparatively higher level of expertise than another appears to be a more reliable local assessment. This transition turns the expertise assessment into a ranking question "is A better than B in context C?" similar to preference ordering [44]. Ferrod et al. [20] turned the problem of detecting the level of expertise of a user from dialogues into a text classification task that concerns and emphasizes expertise in the telecommunication domain. Although their measures are not directly transferable to a general context, the classification methodology confirms that *relative* expertise inferred from classification can avoid defining absolute levels. The literature analysis in this section identified that expertise is highly context-dependent, and that human experience is relevant. To measure the involved expertise in a feedback loop, one does not only need to measure the human experience but also needs to consider the context involved.

3 USER STUDY

We designed the following user study to answer our research questions (**RQ1** and **RQ2**). To understand the impact of expertise on satisfaction, we hypothesize that by using HITL optimization, participants with a higher level of expertise will produce a better outcome quality and perceive higher satisfaction than novice participants. To verify this, we designed a between-group controlled experiment in three problem contexts: text summarization, photo color enhancement, and 3D model simplification. As dependent variables, we measured participants' expertise in a domain context, interactions with the system, and feedback from final questionnaires (individual rating scales and open questions).

3.1 Problem Context

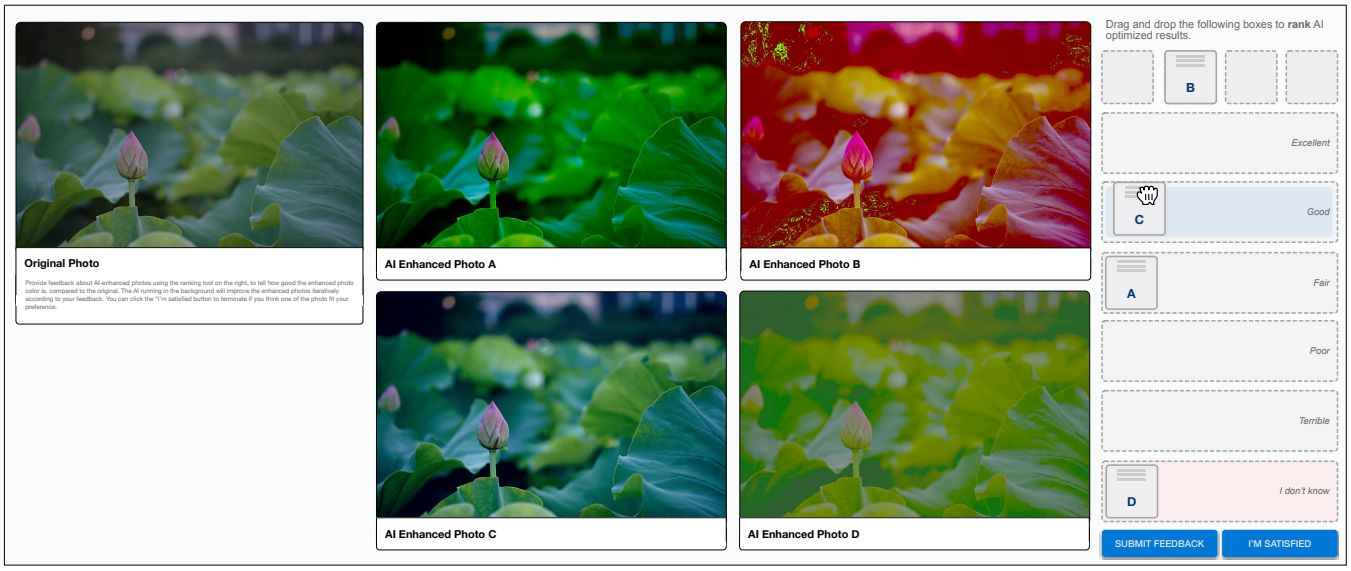
In a HITL optimization context, it is more fitting to use decision-making tasks that sit between pure subjective preference matter (e.g., favorite colors) and well-defined objective optimization problems that can be solved procedurally (e.g., finding the global minimum of a strict convex continuous function). We need to select tasks where users provide ranking feedback using their expertise. A task should also be iterative for observing progress and partially subjective because users could balance the trade-off on different objectives.

We selected tasks that include text summarization, photo color enhancement, and 3D model simplification for the following reasons: 1) They all partially involve rational, objective judgment, and subjective components. 2) Each domain requires different levels of human expertise: text summarization only requires language proficiency, which is a fundamental human expertise; photo color enhancement requires an understanding of aesthetics and color theory; 3D model simplification requires domain-specific technical 3D modeling expertise. 3) All these contexts had been discussed in the HITL optimization literature [39–41, 53, 64] individually but not compared to each other together.

3.2 User Interfaces for Data Collection

Figure 1, and 3 show our UIs in the HITL optimization main task for 3D model simplification, text summarization, and photo color enhancement, respectively. All interfaces collect a participant's expertise at the beginning of the study, then present four variants through the interface. When a task is over, the interface presents six final questions and an open question regarding their satisfaction and overall experience when interacting with the system. In all system interfaces, users can express their ranking choices, and users provide a ranking of the current four result variants on the interface's right side. Additionally, in the 3D model simplification task, a user can rotate, zoom, and move the four models simultaneously to inspect and compare the quality of the models. In line with prior

(a) AI-based text summarization.



(b) AI-based photo color enhancement. The photo is taken from Koyama et al. [40].

Figure 3: The ranking interface for a) text summarization and b) photo color enhancement. In each iteration, the interface presents four options. Participants can drag and drop the top right blocks to a suitable rating region to provide a ranking of the options regarding the given objectives. Each of the regions can contain multiple blocks. Blocks can be put in the "I don't know" region to express an incomplete preference.

work by Ou et al. [53], we use a listwise interface with four variants instead of two pairwise comparisons to increase the collected feedback in each iteration without increasing system processing and data transmission time. After the user submits the ranking information, the background system will utilize this information and then optimize and infer the next optimal set of variants. We also added an "I don't know" container box to the ranking UI and allowed participants to express incomplete preferences. This design is intended to prevent the violation of the completeness axiom.

3.3 Apparatus

We developed the frontend UIs using Material UI[2], React DnD[3], and three.js[4]. Apart from the frontend, our backend *core service* is written in Go[5] for easier concurrency management. It serves our frontend interfaces, data collection, and communications with

[2] https://mui.com/, *last accessed February 13, 2023*
[3] https://react-dnd.github.io/react-dnd/about, *last accessed February 13, 2023*
[4] https://threejs.org, *last accessed February 13, 2023*
[5] https://go.dev, *last accessed February 13, 2023*

other dedicated computing microservices, including *domain services* and an *optimizer service*. The logged data were directly managed using the OS file system with naming conventions. We deployed all services on our institute infrastructure (Ubuntu 20.04, 8-Core Intel Core i9-9900K, 64GB RAM, and NVIDIA GeForce RTX 2080 Ti with 11GB of GPU memory).

3.3.1 Domain services. We implemented three separate domain services. To perform text summarization, we picked the pre-trained BART model via HuggingFace[6]. We implemented an isolated text summarization server using Flask[7] with GPU acceleration. We use Nucleus sampling [29] as a stochastic text decoding strategy[8] for our inference use case because it allows for a bounded hyper-parameter space (between 0 and 1) and can generate diverse human-like sentences in the inference phase. Since we designed our user task to consider the length of summarization as a decision criterion, we used a summarization ratio as a hard limit that controls the text generation length and a length penalty as a selected soft limit that encourages the model to generate shorter text. As a result, our hosted text summarization service allows four adjustable system hyper-parameters at every model inference stage: 1) *summarization ratio*, 2) *length penalty*, 3) *top-p*[8], and 4) *temperature*[8].

For photo color enhancement, we used a parameterized photo enhancer [39–41] as an image processing service for better integration to the core service. This service allows five adjustable system hyper-parameters that are bounded between 0 and 1: 1) *brightness*, 2) *contrast*, 3) *saturation*, 4) *temperature*, and 5) *tint*. Lastly, we used a 3D mesh reducer [53] as a 3D mesh processing service, and it also contains five bounded system hyper-parameters: 1) *simplification ratio*, 2) *border preservation*, 3) *hard edge preservation*, 4) *sharpness preservation*, and 5) *quadrilateral preservation*.

3.3.2 Optimizer service. We implemented the underlying optimizer using BoTorch [3] as a command line service, which reads the user responses to estimate the next optimal system hyper-parameters for exploration. BoTorch provides the EUBO [33] optimizer as one of the state-of-the-art PBO optimizers that consider noisy inputs to estimate system hyper-parameters for pairwise comparisons. We extended EUBO to utilize ranking comparisons to fit a Gaussian process using the user's rank data first. Then, we used the learned latent utility value to fit another Gaussian process and infer the next batch of exploration positions.

3.4 Procedure

The overall procedure is visualized in Figure 4. Participants started the study with an informed consent form and answered initial demographic questions, including their age, gender, and domain expertise. The UI presents a set of evaluation options to participants. The main task is to provide feedback, using the UI, to the AI to 1) for a news article: *summarize the given news articles as much as possible while preserving the meaning of the article*; 2) for a photo: *improve and enhance the color of the photo*; and 3) for a 3D model: *simplify the number of polygons as much as possible while preserving the overall appearance*. Because the optimizers need initialization samples, in the first 4 iterations, a participant evaluated outcomes produced by quasi-random Sobol sampled [56, 69] system hyper-parameters. After acquiring these preference priors from participants, starting from the 5th iteration, the optimizer is used, and participants can freely terminate if satisfied with the current text summary. The task automatically ends after 20 iterations to limit participation time. After termination, participants answered six questions regarding their satisfaction with the system outcomes and their experience using the interface to give feedback. Each participant completed 3x3 Latin square-shuffled news articles[9], photos[10], and 3D models[11]. Example iteratively optimized outcomes are shown in Figure 5.

3.5 Participants

We recruited participants worldwide on Prolific. Because participants had different median completion times in different experimental conditions, we paid between £3 to £9 upon completion, corresponding to an hourly wage of £9/h ($10.4/h). Participants gave informed consent at the beginning of the study; thus, the study adhered to European privacy laws (GDPR). In total, we collected data from 91 participants from 13 countries. To guarantee high-quality responses, we only consider participants if they: 1) had an approval rate of 95%, 2) completed the study only once, 3) answered with consistent demographics, e.g., not more than five years of age difference in the study compared to the platform registration information, and 4) provided their response in at least a reasonable amount of time, i.e., spent longer than 3 seconds in each iteration to read the summarized text and interact with the interface according to our pilot study observations. Therefore, we will report our results based on 60 participants (31 female, 29 male, and no diverse; age $\mu = 26.92$, $\sigma = 6.44$, range 19-52). Each domain context includes 20 participants.

3.6 Measurements

During the study, we measured participants' expertise, their interaction behavior with our developed system, subjective ranking feedback to the system outcomes, objective quality of system outcome, and their subjective satisfaction and open questions regarding their experience.

3.6.1 Expertise Measures. As discussed in Section 2.3, since user expertise is measured differently in prior research, we use a similar approach as Ferrod et al. [20], Ooge and Verbert [52], and combine the following established metrics: 1) self-indication, 2) accumulated work experience, and 3) familiarity with domain problems. For the text context, we ask for their language proficiency, and for the image and mesh contexts, we ask for their self-indicated photo editing and 3D modeling expertise. All contexts asked for their months of work experience as well as when was their last time of experience. Based on the collected data, relative levels of expertise are used in our context for the discussion of expertise, and we normalized these measures among all participants, then grouped participants

[6] https://huggingface.co/sshleifer/distilbart-cnn-12-6, *last accessed February 13, 2023*
[7] https://flask.palletsprojects.com/en/2.2.x/, *last accessed February 13, 2023*
[8] https://huggingface.co/blog/how-to-generate, *last accessed February 13, 2023*
[9] Selected from the CNN daily mail dataset, article IDs: ea06fd0b25cb9793397a, 35f0e33de7923036a97a, 42c027e4ff9730fbb3de. See https://huggingface.co/datasets/cnn_dailymail, *last accessed February 13, 2023*
[10] Selected from Koyama et al. [40]. See https://github.com/yuki-koyama/sequential-gallery/tree/main/resources/scaled, *last accessed February 13, 2023*
[11] Selected model name: stanford bunny, suzanne, and fandisk. See https://github.com/alecjacobson/common-3d-test-models, *last accessed February 13, 2023*

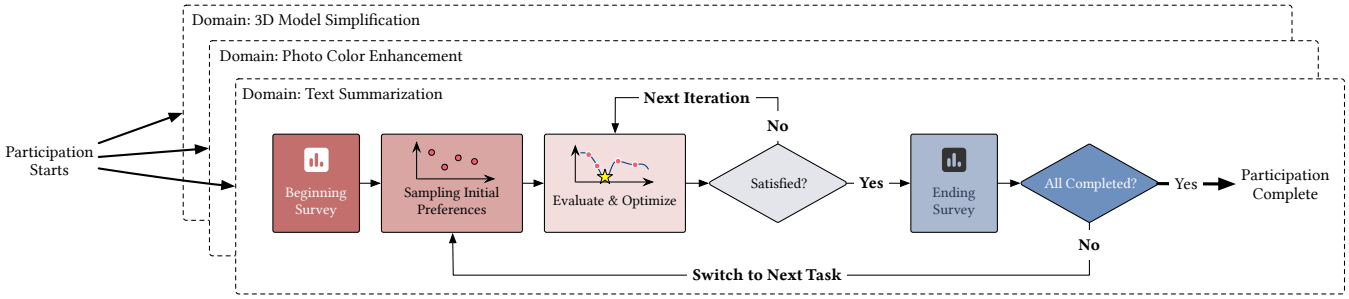


Figure 4: The study procedure: Each participant did one of the three domain contexts (news articles, photos, and 3D models).

into three groups using quantile-based discretization: *novice*, *intermediate*, and *experienced*. Note that the descriptions only represent the relative levels among our recruited participants. In a larger user group, they may be reconsidered as novice or intermediate accordingly.

3.6.2 Behavior Measures. We measured how participants interacted with our developed systems, including 1) decision time: the period between the appearance of the evaluation options and a full ranking is submitted, 2) number of iterations per task, 3) number of incomplete preferences per iteration, 4) number of indifference preferences per task, and 5) number of drag and drop operations to rank given options.

3.6.3 Objective Outcome Quality Measures. In the text summarization context, we measured the total number of words in the outcome texts to validate if participants made progress on the given objectives. We also computed the BLEU [54] and ROUGE [47] (including ROUGE-1, ROUGE-2, and ROUGE-L) scores to measure the objective quality of summarized texts as they are frequently used to evaluate the quality of summarization, and correlate positively with human evaluation. For the photo color enhancement context, we converted the outcome photo from RGB color space to HSV and YUV to directly reflect the relevant optimization metrics. Namely, we computed saturation, contrast, and brightness using a mean of pixel-wise subtraction between source and outcome in the H, S, and V channels correspondingly. Furthermore, we computed the mean pixel-wise difference of U channels in YUV color space for tint changes and similarly in the V channel for temperature changes.

The task of 3D model simplification concerns simplification ratio [8] and perceivable changes regarding visual quality, wireframe quality, and surface quality. The visual quality and wireframe quality are useful indicators concerning human perceptual judgments. In contrast, surface quality is defined at a technical level and was found to be more difficult to perceive visually compared to the other qualities [16]. Therefore, we computed the simplification ratio to validate whether participants progressed on the given objectives. In terms of visual quality, one can use rendering quality from multiple camera views to measure visual quality during mesh simplification. A 3D model can be rendered as a series of images from different perspectives with given rendering settings, such as specified light conditions, camera settings, and rendering algorithms. We use an equally weighted combination of *peak signal-to-noise ratio* [37] (PSNR) and *structural similarity* [68] (SSIM) to measure the rendering difference. For wireframe quality, we computed *scaled Jacobian cell quality* [36] because it was previously suggested to better correlate with the human judgment [16]. The scaled Jacobian cell quantity itself measures how a given face is regularized. Lastly, we sampled two point clouds on the source mesh and the outcome, then used *Chamfer distance* [5] to indicate the surface quality. Surface quality is less observable compared to the other three objectives.

3.6.4 Subjective Measures. After participants ended a task, we asked six questions: Q1) participants' overall subjective satisfaction with the final results; Q2) their confidence if they think they can do a better version by themselves than the system, which was optimized based on their provided feedback; Q3) whether the final outcome matched their expected result; Q4) whether they felt improvements of the result from iteration to iteration; Q5) whether they felt the "I don't know" option was useful, and Q6) whether they believed they gave clear feedback to the AI. We measured these questions using a bipolar slider-based Likert scale. Among these questions, Q1 to Q4 are intended to measure subjective satisfaction.

4 RESULTS

Based on the behavior and subjective questionnaires, we first verify the relationship between user expertise and satisfaction. Then, we evaluate the optimization loops in different domains based on the collected ranking data and objective quality measurements of outcomes.

4.1 Behavior and Subjective Satisfaction

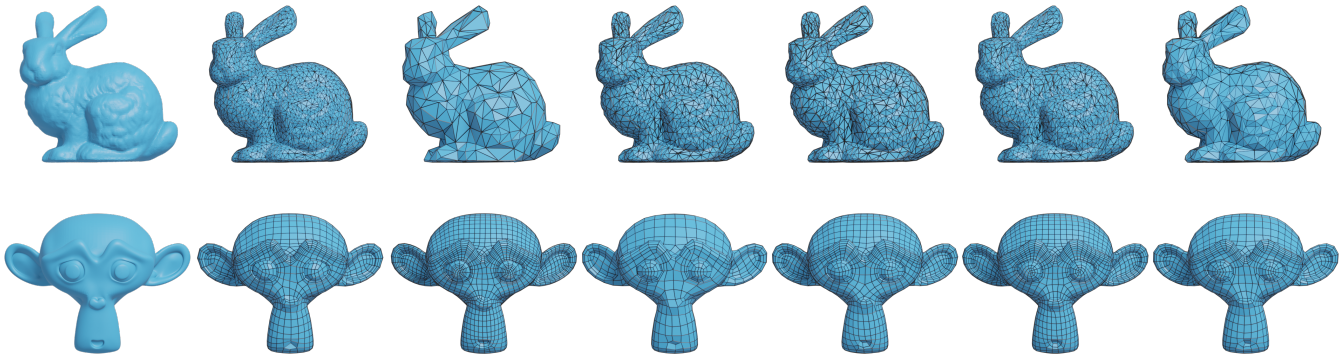
To analyze the behavior and subjective satisfaction, we first group our participants using quantile-based discretization to guarantee each grouped expertise level has an evenly distributed number of participants. Then we assert the data's normality using the Shapiro-Wilk test. We use a t-test to compare the difference between novices and experienced participants for normally distributed data. Otherwise, we report a Wilcoxon rank sum test as a non-parametric approach to compare the differences between novice and experienced participants for measured dependent variables. Full results are available in Section 7.

4.1.1 Inferred Expertise. Our participants reported varied experiences in different domains. They self-indicated English proficiency

Barcelona beat Atletico Madrid 3-0 to remain in touch with Real Madrid in La Liga. Lionel Messi and Deco score for Barca in Barca's fourth straight league win against big-spending rivals. Real keep pace at top of table after second straight league victory at Recreativo H	Barcelona beat Atletico Madrid 3-0 to stay in touch with Primera Liga leaders Real Madrid. Real beat Recreativo Huleva 2-0 and Real Madrid beat Real 2-1 to stay two points clear of Real. Real's first-half goalscorer Gonzalo Higuain scores in the dying minutes to keep Real two points behind Real	Barcelona beat Atletico Madrid 3-0 to stay in touch with Primera Liga leaders Real Madrid. Real beat Recreativo Huleva 2-0 and Real Madrid beat Real 2-1 to stay two points clear of Real. Real's first-half goalscorer Gonzalo Higuain scores in the dying minutes to keep Real two points behind Real	Barcelona beat Atletico Madrid 3-0 to stay in touch with Primera Liga leaders Real Madrid. Lionel Messi scores sixth successive goal of the season as Barcelona beat big-spending Atletico. Real Madrid beat Recreativo Huleva 2-0 in La Liga to keep Real two points clear	Barcelona beat Atletico Madrid 3-0 to stay in touch with Primera Liga leaders Real Madrid. Lionel Messi scores sixth successive goal of the season as Barcelona win 4th straight league game. Real Madrid beat Recreativo Huleva 2-0 and Gonzalo Higuain scored in the dying minutes. Real have made their best start since 1991 but coach Bernd Schuster's rotation policy questioned.

(a) AI-based text summarization.

(b) AI-based photo color enhancement. Original photos are taken from Koyama et al. [40]



(c) AI-based 3D model simplification.

Figure 5: Example outcome sequences from the a) text summarization, b) photo color enhancement, and c) 3D model simplification. From left to right, it shows how the objective was optimized progressively until the final satisfying outcome (far right).

on the CEFR scale[12]: B1 10.00%, B2 30.00%, C1 35.00%, and C2 25.00%. For self-indicated expertise in photo editing: none 25.00%, novice 45.00%, intermediate 25.00%, experienced 5.00%, experts 0.00%. For self-indicated expertise in 3D modeling: none 35.00%, novice 45.00%, intermediate 15.00%, experienced 5.00%, and none indicated themselves as experts.

Participants indicated their period of work experience. For text summarization: No work experience 25%, less than one year of experience 30%, 1 to 5 years 25%, more than 5 years 20%; for photo editing: No work experience 50%, less than one year of experience 10%, 1 to 5 years 30%, more than 5 years 10%; for 3D modeling: No work experience 60%, less than one year of experience 40%. Regarding the recent experience in these domains, for text summarization: Never 5%, in recent 2 weeks 20.0%, 2 weeks to 3 months ago 25.0%, 3 to 6 months ago 10.0%, 6 to 12 months ago 20.0%, 13 to 36 months ago 5.0%, more than 36 months ago 15.0%. for photo editing: Never

[12]CEFR scale: https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale, *last accessed February 13, 2023*

10.0%, in recent 2 weeks 40.0%, 2 weeks to 3 months ago 30.0%, 3 to 6 months ago 5.0%, 6 to 12 months ago 10.0%, 13 to 36 months ago 5.0%; and for 3D modeling: Never 40.0%, in recent 2 weeks 15.0%, 2 weeks to 3 months ago 15.0%, 3 to 6 months ago 5.0%, 6 to 12 months ago 5.0%, 13 to 36 months ago 5.0%, more than 36 months ago 15.0%.

In total, using quantile-based discretization, we inferred participants' level of expertise in the three contexts: text summarization (Novice: 7, Intermediate: 7, Experienced: 6); photo color enhancement (Novice: 7, Intermediate: 7, Experienced: 6); 3D model simplification (Novice: 7, Intermediate: 6, Experienced: 7).

4.1.2 Interaction Behaviors. All measured interaction behavior indicators are visualized in Figure 6. In terms of the decision time, we found a significant difference between novices and experienced participants both in text summarization (W = 8273.00, p = .051; r = -0.14, $CI_{95\%}$=[-0.27, -0.0006]), photo color enhancement (W = 22320.00, p = .041; r = 0.12, $CI_{95\%}$=[0.006, 0.23]), and 3D model simplification (W = 20999.50, p < .001; r = 0.45, $CI_{95\%}$=[0.34, 0.54]). This means *experienced participants are either more thoughtful (e.g., in the text summarization domain) or more effective (e.g., photo color enhancement and 3D simplification) in forming their decision.* For the number of involved iterations, we did not find a significant difference between novices and experienced participants in text summarization (W = 223.00, p = .953; r = 0.01, $CI_{95\%}$=[-0.33, 0.35]) and 3D model simplification (W = 199.50, p = .751; r = 0.06, $CI_{95\%}$=[-0.30, 0.40]). However, we found significant more iteration in photo color enhancement (W = 99.00, p = .008; r = -0.48, $CI_{95\%}$=[-0.71, -0.15]) for experienced participants than novices. The results suggest that *experienced participants explore the solution space significantly more when the feedback loop is more efficient.*

When checking the expressed number of incomplete preferences, we found experienced participants rarely express an incomplete preference, and novices in the 3D model simplification domain express incomplete preference significantly more than experienced participants (W = 249.00, p = .023; r = 0.32, $CI_{95\%}$=[-0.04, 0.60]) domains. However, we did not find a significant difference in text summarization (W = 274.50, p = .081; r = 0.24, $CI_{95\%}$=[-0.10, 0.54]) and in photo color enhancement (W = 144.50, p = .154; r = -0.24, $CI_{95\%}$=[-0.54, 0.13]) contexts. Similarly, we found experienced participants indicated indifference preference significantly more than novices in the photo color enhancement domain (W = 102.50, p = .015; r = -0.46, $CI_{95\%}$=[-0.70, -0.13]) but neither in the text domain (W = 265.00, p = .255; r = 0.20, $CI_{95\%}$=[-0.15, 0.51]) nor the 3D model domain (W = 230.50, p = .242; r = 0.22, $CI_{95\%}$=[-0.14, 0.53]). Regarding the number of ranking interactions to express the preference in an iteration, we found experienced participants express significantly more than novices in text summarization (W = 8439.50, p = .062; r = -0.12, $CI_{95\%}$=[-0.25, 0.02]) but not in photo (W = 19405.50, p = .590; r = -0.03, $CI_{95\%}$=[-0.14, 0.09]) and 3D model (W = 14994.50, p = .516; r = 0.03, $CI_{95\%}$=[-0.09, 0.16]) domains. These results show that *experienced participants express their ranking preference more clearly.* In contrast, *novices might not know if the machine outcome may not be good enough for them, resulting in more incomplete and fewer indifferent preferences.*

4.1.3 Subjective Satisfaction. As mentioned in Section 3.6.4, we measured subjective satisfaction at the end of every task, and from Q1 to Q4, are used to measure the satisfaction. Since Cronbach's α is fairly high α=0.721, $CI_{95\%}$=[0.648, 0.782] in our collected data, we aggregate these questions as satisfaction indicators. See Figure 7.

We conducted an ART ANOVA [70], as the Shapiro-Wilk normality test showed that the data are not normally distributed (W=.964, p<.001). This analysis revealed that *the overall satisfaction of the final system outcome is significantly influenced by the involved expertise* ($F_{2,51}$=7.56, p=.001, η^2=0.23) as well as by the domain context ($F_{2,51}$=3.84, p=.027, η^2=0.13). Moreover, no interaction effect was found ($F_{4,51}$=0.50, p=.733, η^2=0.04).

4.2 Interactions within the Optimization Loop

We analyze three aspects to quantify the overall optimization loop: 1) The directly measured preference utility, i.e., ranking data, from participants. 2) The learned latent utility of the underlying BO optimizer, and 3) The system outcome quality based on objective metrics. For the directly measured preference utility, a higher value of utility represents participants considering the outcome quality is better in the current evaluating options. The learned latent utility represents how the underlying algorithms consider the human is satisfied with the current results based on the ranking responses; a higher value represents BO optimizer considers more satisfaction on the human side. Lastly, the objectively measured outcome quality metrics measure how different an outcome is from the original task input.

4.2.1 Measured and Learned Preference Ranking Utility. As shown in Figure 8, for directly measured preference utility from ranking data, we fitted a linear mixed model [6, 43] (estimated using REML and nloptwrap optimizer) to predict preference utility with involved expertise and exploration iterations. The model included participants as a random effect. Comparing to novice participants ($CI_{95\%}$=[0.49, 0.56], t(3592) = 28.56, p < .001), we found that in all domain contexts, the submitted preference utility from experienced participants is statistically non-significant and negative (β = -0.02, $CI_{95\%}$=[-0.07, 0.03], t(3592) = -0.69, p = .489). The effect of iteration is statistically significant and positive (β = .002, $CI_{95\%}$=[.001, .003], t(3592) = 3.32, p < .001). This means that regardless of the involved expertise, participants behave consistently, and in later iterations, the final ranking utility is higher than at the beginning of HITL optimization.

Regarding the learned latent utility from the BO optimizer, as illustrated in Figure 9, we fitted another linear mixed model (estimated using REML and nloptwrap optimizer) to predict the learned latent utility with involved expertise and exploration iterations. Comparing to novice participants ($CI_{95\%}$=[0.42, 0.46], t(3592) = 42.97, p < .001), the effect of experienced participants is statistically significant and positive (β = 0.03, $CI_{95\%}$=[0.001, 0.06], t(3592) = 2.03, p = .042). But the effect of iteration is statistically non-significant and positive (β = 0.001, $CI_{95\%}$=[-0.0004, 0.002], t(3592) = 1.32, p = .186). This result means that the provided ranking data from experienced participants are more effective and consistent for the BO optimizer than the ranking data from novices.

4.2.2 Objective Outcome Quality. We normalized the iteration sequence and visualized the exploration progress in Figure 10. For analyzing the progress, we fitted linear mixed models for all metrics

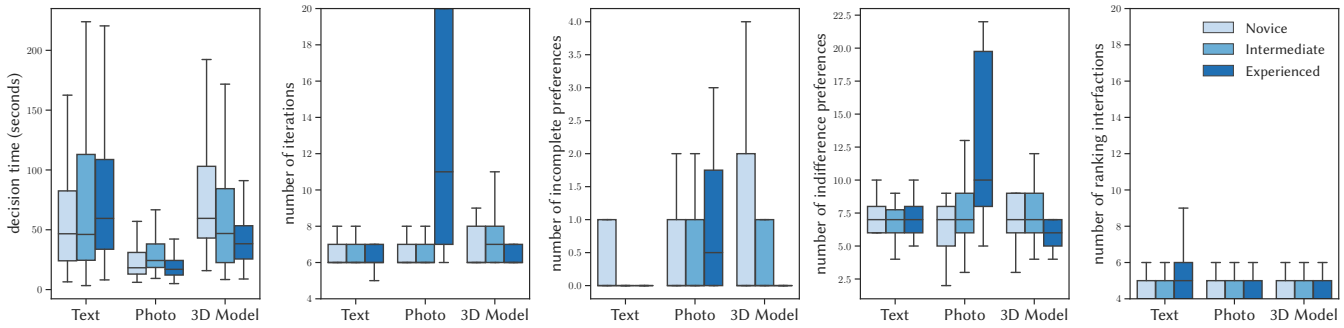


Figure 6: Measured interactions of participants in different domain contexts. Measurements are grouped by the level of expertise. The results indicate that experienced participants express their preferential ranking decisions more clearly than novices. For example, they behave faster in decision time with more iterations or decide slower with more ranking interactions (thoughtful indecision); they also express fewer incomplete and more indifferent preferences.

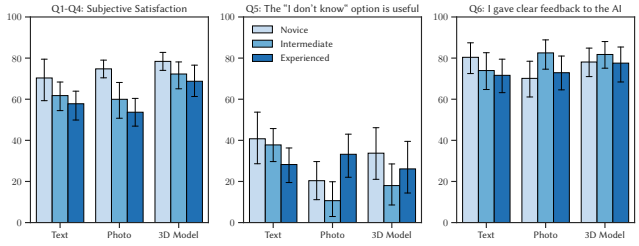


Figure 7: Measured subjective satisfaction, the usefulness of providing incomplete preference option while doing the ranking evaluation. The results suggest that subjective satisfaction significantly decreases when comparing novice and experienced participants. All participants considered allowing expressing incomplete preference less useful, and they gave clear feedback to the AI.

in the text summarization domain. For example, for length metric: comparing to the results produced by novices ($CI_{95\%}$=[51.20, 54.43], t(1192) = 64.14) is as good as the outcome produced by experienced participants (β = -0.20, $CI_{95\%}$=[-2.48, 2.08], t(1192) = -0.17, p = .864), and there are no effects on the involved iteration (β = -0.12, $CI_{95\%}$=[-0.26, 0.03], t(1192) = -1.55, p = .120). These results hold the same as for other metrics. In summary, when comparing to outcomes produced when engaging with novices, the effects of involving experienced participants were statistically non-significant, and the effect of iteration was statistically non-significant and negative. This means novices achieved the same level of performance as experienced participants did. These results hold for all metrics we used for measuring outcome quality.

In the photo color enhancement, except for the contracts (β = -1.45, $CI_{95\%}$=[-2.05, -0.84], t(1192) = -4.71, p < .001) and temperature (β = 0.19, $CI_{95\%}$=[0.004, 0.37], t(1192) = 2.00, p = .045) which are significantly influenced regarding exploration iterations. The effect on brightness using experienced participants is statistically non-significant and positive (β = 0.76, $CI_{95\%}$=[-10.38, 11.89], t(1192) = 0.13, p = .894) and the effect of iteration is statistically non-significant and negative (β = -0.32, $CI_{95\%}$=[-0.78, 0.14], t(1192) = -1.37, p = .172), when compared to novices ($CI_{95\%}$=[-7.11, 8.02], t(1192) = 0.12, p = .907), and these results are the same for saturation and tint metrics.

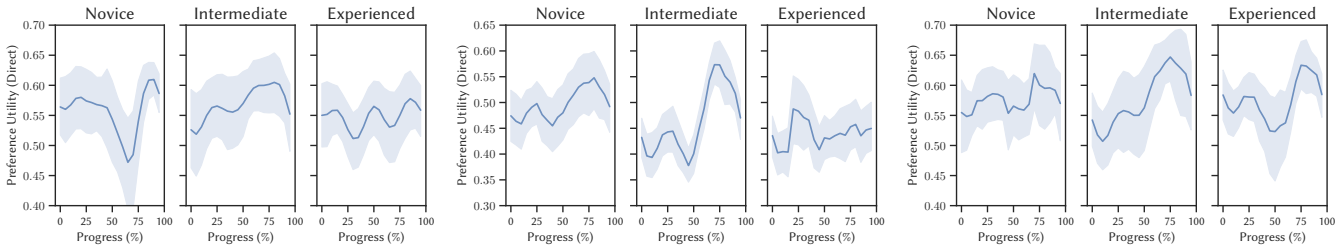
Lastly, for 3D model simplification, we found that experienced participants (β = 0.003, $CI_{95\%}$=[0.0001, 0.007], t(1192) = 2.00, p = .046) outperformed novices ($CI_{95\%}$=[-0.002, 0.002], t(1192) = 0.10) only in keeping surface distance low, meaning better in maintaining surface quality. We did not find significant differences in other metrics when comparing experienced users' and novices' outcomes. This result means that experienced participants are better at identifying technical differences as surface quality is less observable, as discussed in Section 3.6.3. However, novices can achieve expert-level performance under the HITL optimization context, similar to other contexts.

5 DISCUSSION AND IMPLICATIONS

The results in Section 4 could be summarized into two major observations: 1) Novices can achieve expert-level performance in objective quality in all cases. 2) Participants with higher expertise show more explicit preferences, dissatisfaction, and iterations, but novices are more quickly terminated and show more satisfaction. Below, we will discuss what implications we think these observations might have.

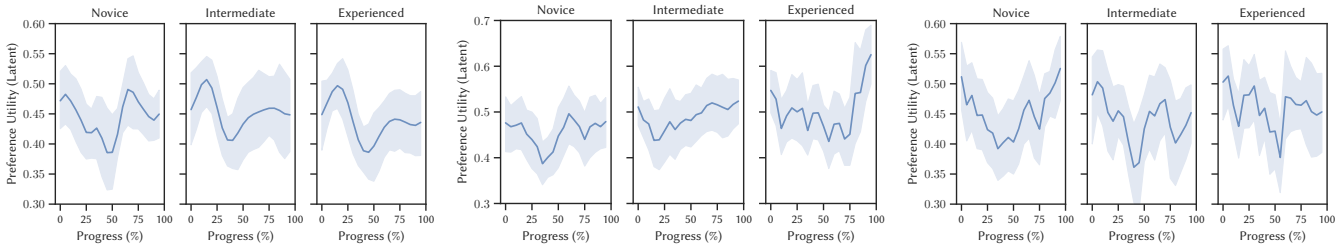
5.1 Outcome Quality and Pareto Optimality

When we have a well-defined metric that can measure the quality of an outcome, the optimization process could be done procedurally using a machine alone. However, in reality, the outcome quality is often characterized by a set of metrics, and *Pareto optimality* [55] is a useful concept for discussing machine rationality regarding its outcome quality. *Pareto optimality* describes a trade-off situation where a system outcome is optimal if any improvements in one objective result in the deterioration of others. This trade-off is also called the *Pareto front*, and outcomes on this front refer to *Pareto frontiers*. Conceptually, the Pareto optimality captures the measurable components when evaluating an outcome, whereas non-measurable parts reflect more about the subjective matter. Let $\mathcal{P}_s$ be the system parameter space defined by $[0, 1]^r (s \in \mathrm{N}^+)$, and O be the outcome



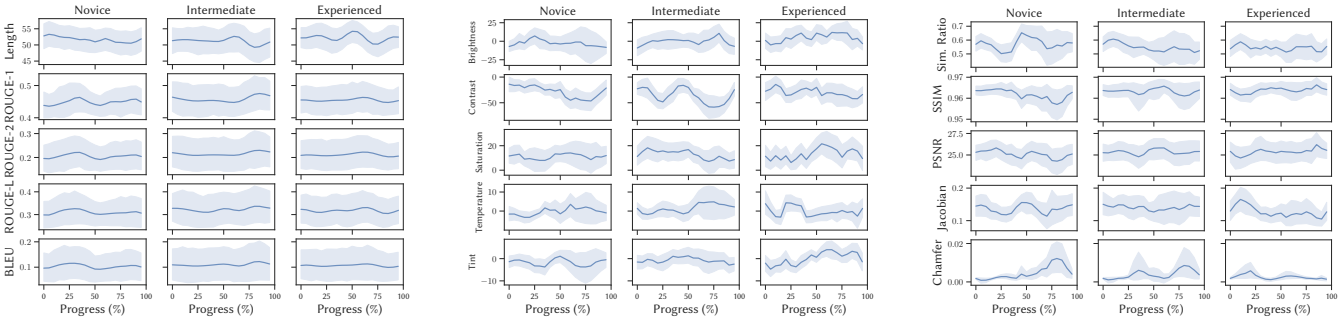
(a) Text Summarization (b) Photo Color Enhancement (c) 3D Model Simplification

Figure 8: Directly measured preference utility: The utility values are normalized from rating labels (Terrible to Excellent). The results indicate that regardless of the involved expertise, participants behave consistently, and in later iterations, the final ranking utility is higher than at the beginning of HITL optimization



(a) Text Summarization (b) Photo Color Enhancement (c) 3D Model Simplification

Figure 9: Learned latent preference utility: The inferred utility values from the machine side (i.e., Bayesian optimization). Our results indicate that provided ranking data from experienced participants are more consistent and more effective in the learning process for the BO optimizer to learn than the ranking data from novices.



(a) Text Summarization (b) Photo Color Enhancement (c) 3D Model Simplification

Figure 10: Objective Quality of System Outcomes: Each context used five metrics to measure the outcome. Experts can identify the technical difference compared to novices, such as minimizing Chamfer distance in 3D model simplification. Overall, our results indicate that novices produce expert-level performance in objective quality.

space generated from the parameter space. Then, the rational component of a HITL optimization is to explore the outcome space O concerning a given set of objective metrics $\mathcal{M}_t (t \in \mathbb{N}^+)$. The Pareto front $\mathcal{F}$ is determined by the outcome space and specified metrics, which essentially depends on the parameter space and metrics, i.e., $\mathcal{F}(\mathcal{P}_s, \mathcal{M}_t)$, which captures the boundary of machine rationality and HITL optimization could be considered as the exploration in this space to reach the Pareto front.

This concept avoids the aggregation problem of contradicted multi-objective objectives, such as in our user tasks, participants need to summarize the text while preserving the meaning or simplify 3D models as much as possible while keeping the overall appearance. However, note that converging to the true Pareto optimal set has a technical challenge, and yet still in active research [17, 59], as there might be an infinite amount of candidates, and metrics might interact with each other. Instead of evaluating whether an

outcome is a Pareto frontier, it is more useful to discuss whether the optimization made any progress to guarantee the final outcome is more dominant than the initial ones.

In our results, we showed that both novices and experienced participants improved the objective measures and could achieve a similar level of quality, meaning the final outcomes are Pareto dominant than the initial ones. Under the Pareto optimality framework, the BO learns the underlying preference using users' ranking choices, which tend to converge to different non-Pareto optimal results. But since the BO optimizer assumes human has a stable preference utility function that will eventually converge, we argue that novice participants do not have enough evaluation metrics in mind, and the system outcome does not necessarily need to arrive at the front. In contrast, experts attempt to keep optimizing or exploring other objectives when machine rationality already reaches the objective Pareto front. Hence, compared to experienced participants who potentially evaluate more metrics than the machine, more flaws might be discovered in this process, and cause either more uncertain in expressing its decision and causing more decision time (e.g., in text summarization) or easier to form a decision and cause less decision time (e.g., in the photo and 3D model contexts). Since experts report significantly higher dissatisfaction than novices, we argue that this result shows a mismatch of the Pareto front between the participants and machine rationality, and the source of the dissatisfaction comes from the involved expertise.

5.2 Expertise and Satisficing Decision Strategy

Based on the analysis of the outcome quality from the HITL optimization loop, we did not find enough evidence to indicate a significant difference regarding the quality of the system outcome between different levels of expertise. However, with increasing expertise, overall user satisfaction decreases, and the number of iterations increases. This observed behavior matches the maximizing decision strategy since participants are asked to terminate at satisfaction, and experts attempt to explore the solution space significantly more than novices. Since the involved expertise is increased, more flaws in the system may be discovered in this process, resulting in more dissatisfaction. This observation suggests that we could involve more expertise to identify more system flaws iteratively while exploring the solution space. Although machine rationality would not be improved without a reparameterization of the underlying algorithm, this observed behavior could be used as an indicator in hindsight analysis to inform system designers to 1) improve underlying machine rationality, 2) further improve the HITL optimization process, and 3) better support users to explore desired solutions. For novices, using a satisficing strategy is good enough to get to expert-level performance with the help of HITL optimization.

5.3 The Impact of Involved Expertise

The objective outcome quality might not depend on the involved human expertise when a machine learner baked enough domain knowledge in its underlying algorithm. What might be the "minimum" required expertise to obtain meaningful machine outputs, then? What if a user constantly provides flawed random choices? Intuitively, such a condition would not benefit a preference-optimized HITL system. Admittedly, to evaluate the behavior between "zero expertise" and "novice," we could program a random choice generator to test and observe the results. Still, we are bound to a limited observation time and two implicit assumptions. The first assumption is that the expertise level has a total order, and a random choice generator is a minimum element for all levels of expertise. Second, a random choice generator can never produce a meaningful outcome in the context of HITL optimization.

These two assumptions might be considered true at first sight. However, we cannot compare the amount of expertise from a random choice generator or an intelligent human being. Notably, the Borel–Cantelli lemma [14][13], states that with an infinite number of events, the probability[14] of observing a meaningful result is 1. This theory explains that even with a random choice generator, as long as it continues to generate choices, a meaningful sequence of choices eventually will occur, such that the HITL system can produce desired outcomes. In other words, this theoretical fact endorses that a sufficient amount of expertise could be beneficial to produce meaningful outcomes in a short amount of time comparably, and our results complement that more involved expertise creates increased iterations of interactions for explorations.

5.4 Limitations and Future Work

Although we allowed users to express "I don't know" as their incomplete preference, a participant may still provide a sub-optimal ranking due to fatigue from a long time of participation or other relevant reasons, resulting in the violation of the incompleteness assumption. From an algorithmic perspective, although the PBO handles ideal randomized choices, the provided ranking choices might even be worse than assumed Gaussian distributed random choices due to subjective reasons. Besides, the underlying preferences might change at every iteration. For example, experts may further reason for using the system outcomes or trying to make sense of the sequential outcomes. Instead, novice users judge locally, making their behavior much more stable. The choice of objective quality evaluation metrics may also impact the interpretation of the optimization process due to their interaction effect.

One of the conventional motivations for developing an objective metric is to use it to predict human judgments. The development of an objective metric implicitly assumes common sense among the crowd, and the metric may not be suitable for measuring individual preferences. Instead of asking users for their judgment to explore the solution space, it might be more interesting for future research to utilize human judgment more in exploring dynamic solution spaces where human is only involved when the machine reaches its boundary of rationality. Furthermore, instead of evaluating the impact of expertise on the exploration behavior of one static solution space, we could evaluate the interaction effect of the involved expertise and the underlying HITL optimizer. For instance, one could design an experiment to understand the decision behavior on the Pareto front where all machine-proposed options are objectively optimal. It would be interesting to check how the involved expertise impacts the decision behavior among all objectively optimal Pareto frontiers and, thus, better understand the difference between subjective and objective Pareto fronts.

[13] In proposition 10.2.2 (b).
[14] Strictly speaking, the event happens *almost surely* as the Lebesgue measure is 1.

6 SUMMARY AND OUTLOOK

In this paper, we evaluated three example contexts to understand the impact of involving different levels of human expertise in HITL optimization on the subjective satisfaction of system outcomes quality. Our study answered our initial research questions: **RQ1** using a human in the loop to optimize system outcomes allows novice users to achieve expert-level performance; **RQ2** with decreasing expertise, the eventual subjective satisfaction increases, and the entire process terminates faster.

Our findings contradict the intuition that using higher expertise leads to better results. Instead, when collaborating with a machine learner, users without a sufficient amount of domain expertise can still show a compatible level of performance as experts, with even higher satisfaction. We argue interpretations of these observations: 1) When humans do not have enough insights to evaluate the quality of the system outcomes, the eventual result reflects the performance of the machine algorithm. 2) Expert users express less satisfaction when having more insights than the underlying algorithm, and the underlying machine rationality limits the outcome of the optimization loop. In this case, the satisfaction of a human can be used as an indicator to inform system designers further to improve their underlying machine algorithm. The insights suggest optimization using human feedback may be more helpful and favor exploration purposes rather than using them for exploiting the solution space covered by machine rationality. Our results bring us closer to better human models and system design principles for exploiting human intelligence. Inferring and adapting to user expertise also play a pivotal role in achieving successful interaction. An intelligent machine system that adequately considers human expertise can help users improve their skills and achieve higher user satisfaction.

7 OPEN SCIENCE

We encourage readers to reproduce and extend our results. We open-sourced the collected dataset, systems, and analysis scripts at https://changkun.de/s/expertise-loop.

ACKNOWLEDGMENTS

We thank Heiko Drewes for his insights regarding expertise; Dennis Dietz, Yi Xia, Guoliang Xue, Yifei Zhan, Daniel Buschek, and Francesco Chiossi for their inspiring discussions regarding user satisfaction, preference, choice, and decision-making; Eyke Hüllermeier and Karlson Pfannschmidt for the useful discussions of context-depended ranking.

REFERENCES

[1] Paul Anand. 1987. Are the preference axioms really rational? *Theory and Decision* 23, 2 (Sep 1987), 189–214. https://doi.org/10.1007/BF00126305

[2] Jukka Arvo, Antti Euranto, Lauri Jarvenpaa, Teijo Lehtonen, and Timo Knuutila. 2015. *3D Mesh Simplification – A survey of algorithms and CAD model simplification tests.* University of Turku, Turku, Finland. http://urn.fi/URN:ISBN:978-951-29-6202-0

[3] Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. 2020. BOTORCH: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1807, 15 pages.

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 0101 (Jul 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[5] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. 1977. Parametric Correspondence and Chamfer Matching: Two New Techniques for Image Matching. https://apps.dtic.mil/sti/citations/ADA458355

[6] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting Linear Mixed-Effects Models using lme4. https://doi.org/10.48550/ARXIV.1406.5823

[7] Alessio Benavoli, Dario Azzimonti, and Dario Piga. 2021. Choice functions based multi-objective Bayesian optimisation. https://doi.org/10.48550/arXiv.2110.08217

[8] Mario Botsch, Leif Kobbelt, Mark Pauly, Pierre Alliez, and Bruno Lévy. 2010. *Polygon mesh processing*. CRC press, USA.

[9] Lyle Bourne, James Kole, and Alice Healy. 2014. Expertise: defined, described, explained. *Frontiers in Psychology* 5 (2014), 186. https://doi.org/10.3389/fpsyg.2014.00186

[10] Eduard Brandstätter, Gerd Gigerenzer, and Ralph Hertwig. 2006. The Priority Heuristic: Making Choices Without Trade-Offs. *Psychological review* 113, 2 (Apr 2006), 409–432. https://doi.org/10.1037/0033-295X.113.2.409

[11] Eric Brochu, Tyson Brochu, and Nando de Freitas. 2010. A Bayesian Interactive Optimization Approach to Procedural Animation Design. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Madrid, Spain) *(SCA '10)*. Eurographics Association, Goslar, DEU, 103–112. https://dl.acm.org/doi/10.5555/1921427.1921443

[12] Eric Brochu, Nando de Freitas, and Abhijeet Ghosh. 2007. Active preference learning with discrete choice data. In *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07)*. Curran Associates Inc., Red Hook, NY, USA, 409–416. https://dl.acm.org/doi/abs/10.5555/2981562.2981614

[13] Liwei Chan, Yi-Chi Liao, George B. Mo, John J. Dudley, Chun-Lien Cheng, Per Ola Kristensson, and Antti Oulasvirta. 2022. Investigating Positive and Negative Qualities of Human-in-the-Loop Optimization for Designing Interaction Techniques. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3491102.3501850

[14] Donald L. Cohn. 2013. *Probability*. Springer, New York, NY. 307–371 pages. https://doi.org/10.1007/978-1-4614-6956-8_10

[15] Harry Collins. 2013. Three dimensions of expertise. *Phenomenology and the Cognitive Sciences* 12, 2 (June 2013), 253–273. https://doi.org/10.1007/s11097-011-9203-5

[16] Massimiliano Corsini, Mohamed-Chaker Larabi, Guillaume Lavoué, Oldřich Petřík, Libor Váša, and Kai Wang. 2013. Perceptual Metrics for Static and Dynamic Triangle Meshes. *Computer Graphics Forum* 32, 1 (2013), 101–125. https://doi.org/10.1111/cgf.12001

[17] Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. 2021. Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., Red Hook, NY, USA, 2187–2200. https://proceedings.neurips.cc/paper/2021/file/11704817e347269b7254e744b5e22dac-Paper.pdf

[18] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology. General* 144, 1 (Feb 2015), 114–126. https://doi.org/10.1037/xge0000033

[19] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI'21)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3411764.3445188

[20] Roger Ferrod, Federica Cena, Luigi Di Caro, Dario Mana, and Rossana Grazia Simeoni. 2021. Identifying users' domain expertise from dialogues. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 29–34. https://doi.org/10.1145/3450614.3461683

[21] Michael Garland and Paul S. Heckbert. 1998. Simplifying surfaces with color and texture using quadric error metrics. In *Proceedings of the conference on Visualization '98 (VIS '98)*. IEEE Computer Society Press, Washington, DC, USA, 263–269. https://doi.org/10.5555/288216.288280

[22] S.K. Garrett, B.S. Caldwell, E.C. Harris, and M.C. Gonzalez. 2009. Six dimensions of expertise: a more comprehensive definition of cognitive expertise for team coordination. *Theoretical Issues in Ergonomics Science* 10, 2 (March 2009), 93–105. https://doi.org/10.1080/14639220802059190

[23] Gerd Gigerenzer and Henry Brighton. 2009. Homo heuristicus: why biased minds make better inferences. *Topics in Cognitive Science* 1, 1 (Jan 2009), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

[24] Javier González, Zhenwen Dai, Andreas Damianou, and Neil D. Lawrence. 2017. Preferential Bayesian Optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, Australia, 1282–1291. https://doi.org/10.5555/3305381.3305514

[25] William M. Grove and Paul E. Meehl. 1996. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction

procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law* 2, 2 (1996), 293–323. https://doi.org/10.1037/1076-8971.2.2.293
[26] Till Grüne. 2004. The Problems of Testing Preference Axioms with Revealed Preference Theory. *Analyse & Kritik* 26, 2 (Nov 2004), 382–397. https://doi.org/10.1515/auk-2004-0204
[27] Daniel M. Hausman. 2011. *Preference, Value, Choice, and Welfare.* Cambridge University Press, New York, USA.
[28] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User Trust in Intelligent Systems: A Journey Over Time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 164–168. https://doi.org/10.1145/2856767.2856811
[29] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration. https://doi.org/10.48550/ARXIV.1904.09751
[30] Sheena S Iyengar, Rachael E Wells, and Barry Schwartz. 2006. Doing better but feeling worse: Looking for the "best" job undermines satisfaction. *Psychological Science* 17, 2 (2006), 143–150. https://doi.org/10.1111/j.1467-9280.2006.01677.x
[31] Wenzel Jakob, Marco Tarini, Daniele Panozzo, and Olga Sorkine-Hornung. 2015. Instant Field-Aligned Meshes. *ACM Transactions on Graphics* 34, 6 (Oct. 2015), 189:1–189:15. https://doi.org/10.1145/2816795.2818078
[32] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender Systems: An Introduction.* Cambridge University Press, New York, USA.
[33] Zhiyuan Jerry Lin, Raul Astudillo, Peter Frazier, and Eytan Bakshy. 2022. Preference Exploration for Efficient Bayesian Optimization with Multiple Outcomes. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 151)*, Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (Eds.). PMLR, Virtual, 4235–4258. https://proceedings.mlr.press/v151/jerry-lin22a.html
[34] Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise.* Harper-Collins UK, UK.
[35] René F. Kizilcec. 2016. 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 (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2390–2395. https://doi.org/10.1145/2858036.2858402
[36] Patrick M. Knupp. 2000. Achieving finite element mesh quality via optimization of the Jacobian matrix norm and associated quantities. Part I—a framework for surface mesh optimization. *Internat. J. Numer. Methods Engrg.* 48, 3 (2000), 401–420. https://doi.org/10.1002/(SICI)1097-0207(20000530)48:3<401::AID-NME880>3.0.CO;2-D
[37] Jari Korhonen and Junyong You. 2012. Peak signal-to-noise ratio revisited: Is simple beautiful?. In *2012 Fourth International Workshop on Quality of Multimedia Experience*. IEEE, Melbourne, VIC, Australia, 37–38. https://doi.org/10.1109/QoMEX.2012.6263880
[38] Ben Kotzee and Jp Smit. 2018. Two Social Dimensions of Expertise. In *Education and Expertise*. John Wiley & Sons, Ltd, New Jersey, USA, 99–116. https://doi.org/10.1002/9781119527268.ch5
[39] Yuki Koyama, Daisuke Sakamoto, and Takeo Igarashi. 2016. SelPh: Progressive Learning and Support of Manual Photo Color Enhancement. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2520–2532. https://doi.org/10.1145/2858036.2858111
[40] Yuki Koyama, Issei Sato, and Masataka Goto. 2020. Sequential gallery for interactive visual design optimization. *ACM Transactions on Graphics* 39, 4 (July 2020), 88:88:1–88:88:12. https://doi.org/10.1145/3386569.3392444
[41] Yuki Koyama, Issei Sato, Daisuke Sakamoto, and Takeo Igarashi. 2017. Sequential Line Search for Efficient Visual Design Optimization by Crowds. *ACM Trans. Graph.* 36, 4, Article 48 (jul 2017), 11 pages. https://doi.org/10.1145/3072959.3073598
[42] Vijay Krishna and John Morgan. 2001. A Model of Expertise. *The Quarterly Journal of Economics* 116, 2 (May 2001), 747–775. https://doi.org/10.1162/00335530151144159
[43] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 13 (2017), 1–26. https://doi.org/10.18637/jss.v082.i13
[44] Michael D. Lee, Mark Steyvers, Mindy de Young, and Brent Miller. 2012. Inferring Expertise in Knowledge and Prediction Ranking Tasks. *Topics in Cognitive Science* 4, 1 (2012), 151–163. https://doi.org/10.1111/j.1756-8765.2011.01175.x
[45] Thibault Lescoat, Hsueh-Ti Derek Liu, Jean-Marc Thiery, Alec Jacobson, Tamy Boubekeur, and Maks Ovsjanikov. 2020. Spectral Mesh Simplification. *Computer Graphics Forum* 39, 2 (2020), 315–324. https://doi.org/10.1111/cgf.13932
[46] Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. 2019. Constrained Bayesian Optimization with Noisy Experiments. *Bayesian Analysis* 14, 2 (Jun 2019), 495–519. https://doi.org/10.1214/18-BA1110
[47] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013
[48] J. Marks, B. Andalman, P. A. Beardsley, W. Freeman, S. Gibson, J. Hodgins, T. Kang, B. Mirtich, H. Pfister, W. Ruml, K. Ryall, J. Seims, and S. Shieber. 1997. Design Galleries: A General Approach to Setting Parameters for Computer Graphics and Animation. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques (SIGGRAPH '97)*. ACM Press/Addison-Wesley Publishing Co., USA, 389–400. https://doi.org/10.1145/258734.258887
[49] Petrus Mikkola, Milica Todorović, Jari Järvi, Patrick Rinke, and Samuel Kaski. 2020. Projective Preferential Bayesian Optimization. In *International Conference on Machine Learning*. PMLR, MLResearchPress, Online, 6884–6892. https://doi.org/10.5555/3524938.3525577
[50] Arthur I. Miller. 2019. *The Artist in the Machine: The World of AI-Powered Creativity.* MIT Press, Cambridge, Massachusetts, USA.
[51] Robert M. Monarch. 2021. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI.* Simon and Schuster, USA.
[52] Jeroen Ooge and Katrien Verbert. 2021. Trust in Prediction Models: a Mixed-Methods Pilot Study on the Impact of Domain Expertise. https://doi.org/10.48550/arXiv.2109.08183 arXiv:arXiv:2109.08183
[53] Changkun Ou, Daniel Buschek, Sven Mayer, and Andreas Butz. 2022. The Human in the Infinite Loop: A Case Study on Revealing and Explaining Human-AI Interaction Loop Failures. In *Proceedings of Mensch Und Computer 2022 (MuC '22)*. Association for Computing Machinery, Darmstadt, Germany, 1–11. https://doi.org/10.1145/3543758.3543761
[54] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135
[55] Vilfredo Pareto. 1912. Manuel d'économie politique. *Bull. Amer. Math. Soc* 18, 462-474 (1912), 3.
[56] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32* (Beijing, China) *(ICML'14)*. JMLR.org, Beijing, China, II–1278–II–1286. https://proceedings.mlr.press/v32/rezende14.html
[57] Robert P. Rooderkerk, Harald J. Van Heerde, and Tammo H.A. Bijmolt. 2011. Incorporating Context Effects into a Choice Model. *Journal of Marketing Research* 48, 4 (Aug 2011), 767–780. https://doi.org/10.1509/jmkr.48.4.767
[58] David Schmidtz. 2004. Satisficing as a Humanly Rational Strategy. In *Satisficing and Maximizing: Moral Theorists on Practical Reason*, Michael Byron (Ed.). Cambridge University Press, New York, USA, 30–59. https://doi.org/10.1017/CBO9780511617058.003
[59] Adriana Schulz, Harrison Wang, Eitan Grinspun, Justin Solomon, and Wojciech Matusik. 2018. Interactive exploration of design trade-offs. *ACM Transactions on Graphics* 37, 4 (Jul 2018), 131:1–131:14. https://doi.org/10.1145/3197517.3201385
[60] Barry Schwartz, Andrew Ward, John Monterosso, Sonja Lyubomirsky, Katherine White, and Darrin R. Lehman. 2002. Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology* 83, 5 (Nov 2002), 1178–1197. https://doi.org/10.1037/0022-3514.83.5.1178
[61] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. 2016. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* 104, 1 (Jan. 2016), 148–175. https://doi.org/10.1109/JPROC.2015.2494218
[62] Eero Siivola, Akash Kumar Dhaka, Michael Riis Andersen, Javier González, Pablo García Moreno, and Aki Vehtari. 2021. Preferential Batch Bayesian Optimization. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, Gold Coast, Australia, 1–6. https://doi.org/10.1109/MLSP52302.2021.9596494
[63] Herbert A. Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* 69, 1 (Feb. 1955), 99–118. https://doi.org/10.2307/1884852
[64] Edwin Simpson, Yang Gao, and Iryna Gurevych. 2020. Interactive Text Ranking with Bayesian Optimization: A Case Study on Community QA and Summarization. *Transactions of the Association for Computational Linguistics* 8 (12 2020), 759–775. https://doi.org/10.1162/tacl_a_00344
[65] Louis L. Thurstone. 1927. A law of comparative judgment. *Psychological Review* 34, 4 (1927), 273–286. https://doi.org/10.1037/h0070288
[66] Jeffrey W. Treem and Paul M. Leonardi. 2016. *Expertise, Communication, and Organizing*. Oxford University Press, New York, NY, USA.
[67] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (Sept. 1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124
[68] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (Apr 2004), 600–612. https://doi.org/10.1109/TIP.2003.819861
[69] James T. Wilson, Riccardo Moriconi, Frank Hutter, and Marc Peter Deisenroth. 2017. The reparameterization trick for acquisition functions. https://doi.org/10.48550/ARXIV.1712.00424
[70] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova

Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[71] Yonghao Yue, Yuki Koyama, Issei Sato, and Takeo Igarashi. 2021. User interfaces for high-dimensional design problems: from theories to implementations. In *ACM SIGGRAPH 2021 Courses (SIGGRAPH '21)*. Association for Computing Machinery, New York, NY, USA, 1–34. https://doi.org/10.1145/3450508.3464551

[72] Yijun Zhou, Yuki Koyama, Masataka Goto, and Takeo Igarashi. 2021. Interactive Exploration-Exploitation Balancing for Generative Melody Composition. In *26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 43–47. https://doi.org/10.1145/3397481.3450663