The Human in the *Infinite Loop*: A Case Study on Revealing and Explaining Human-AI Interaction Loop Failures

Changkun Ou
LMU Munich
Germany
research@changkun.de

Daniel Buschek
University of Bayreuth
Germany
daniel.buschek@uni-bayreuth.de

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

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

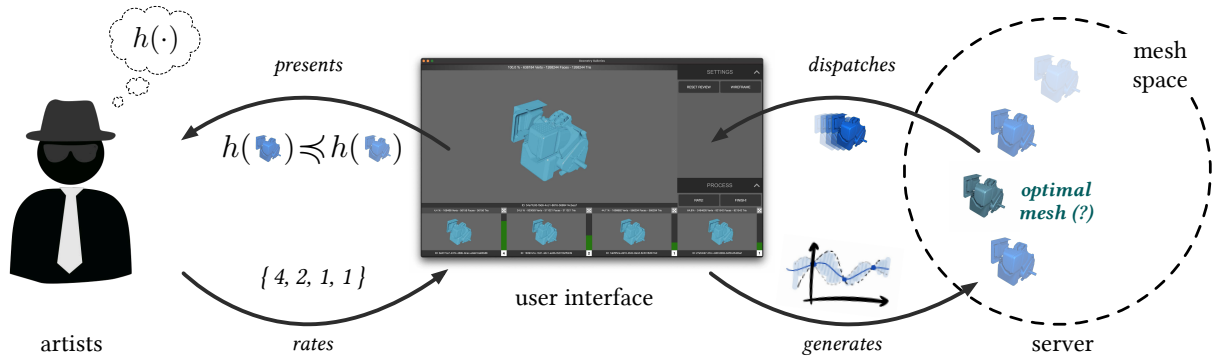


Figure 1: A human-in-the-loop 3D model processing system: A server generates differently processed variations of a complex 3D model and dispatches them to a user interface, which presents those variants to a 3D artist, who in turn rates them. Based on these ratings, new parameter settings are generated and a new set of variations is computed and evaluated again. The process repeats until a satisfactory 3D model is found, that minimizes the number of faces while maintaining as much as possible of its overall appearance.

ABSTRACT

Interactive AI systems increasingly employ a human-in-the-loop strategy. This creates new challenges for the HCI community when designing such systems. We reveal and investigate some of these challenges in a case study with an industry partner, and developed a prototype human-in-the-loop system for preference-guided 3D model processing. Two 3D artists used it in their daily work for 3 months. We found that the human-AI loop often did not converge towards a satisfactory result and designed a lab study (N=20) to investigate this further. We analyze interaction data and user feedback through the lens of theories of human judgment to explain the observed human-in-the-loop failures with two key insights: 1) optimization using preferential choices lacks mechanisms to deal with inconsistent and contradictory human judgments; 2) machine outcomes, in turn, influence future user inputs via heuristic biases and loss aversion. To mitigate these problems, we propose descriptive UI design guidelines. Our case study draws attention to challenging and practically relevant imperfections in human-AI loops that need to be considered when designing human-in-the-loop systems.

CCS CONCEPTS

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

KEYWORDS

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

ACM Reference Format:
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 *Mensch und Computer 2022 (MuC '22), September 4–7, 2022, Darmstadt, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543758.3543761

MuC '22, September 4–7, 2022, Darmstadt, Germany

ACM ISBN 978-1-4503-9690-5/22/09…$15.00
https://doi.org/10.1145/3543758.3543761

1 INTRODUCTION

With the increasing interest in human-AI interaction, *human-in-the-loop (HITL)* [45] systems have been applied to a wide range of domains, such as material design [6], animation design [5], photo color enhancement [39], image restoration [64], and more [11, 21, 38, 67]. These systems actively exploit human choices for optimizing machine results. They propose a set of design alternatives and then iteratively adapt their results based on user preference feedback,

thereby increasing the quality of the system outcomes and the satisfaction of the human involved while simultaneously speeding up the process. The ancestor of these approaches is the Design Galleries approach [42]. Its authors state that "Design Gallery interfaces are a useful tool for many computer graphics applications that require tuning parameters to achieve desired effects". They proposed a generic user interface (UI) and emphasized techniques for dispersing parameter settings, but they implicitly assumed that the process will always converge and end in a desired solution.

Despite these successful examples, obtaining a desirable result in an interactive process remains a challenging and important issue in general, since it is hard to "guarantee", but crucial for such systems to be useful in practice. We argue that to make progress here, it is particularly insightful for HCI researchers to investigate practical deployments in real workflows of domain experts. Concretely, our work is guided by the following two research questions, the first of which first relates to our overarching research interest and the second one to the specific case study reported here:

RQ1 *As a domain-specific example, how can the HITL strategy support users in 3D model processing tasks?*

RQ2 *Which challenges arise for human-AI loops in practice and how can we address them?*

To address these questions, we conducted a case study with an industry partner, in which we applied the HITL strategy to the challenging and practically relevant use case of a 3D modeling workflow: Our prototype system helps 3D artists with the problem of *polygon reduction* [17, 18] that involves tuning control parameters to remove polygons (mainly triangles or quadrilaterals) from 3D models while preserving their overall appearance (see Figure 4a). Although fully algorithmic solutions have improved over the past decades (for instance, using directional fields [25]), they still suffer from problems [1] and often produce unusable results for large models or geometric edge cases. In the meantime, since it often takes 3D artists up to months or years to understand and thus productively work with a new algorithm, this presents a promising target to use an interactive approach in which designers only have to provide their interactive choices to obtain the desired outcome. Thus, we developed a system that optimizes polygon reduction with regard to artists' preferences over the quality of the results based on *Bayesian optimization (BO)* [20, 54]. Figure 1 illustrates its general architecture.

After fine-tuning the system using pilot usage feedback, two professional 3D artists used our system for three months in their real-world 3D workflow. Analyzing this deployment and its usage logs, we found a high failure rate when using a human in the loop. Crucially, final ratings from artists diverged from partially inconsistent or inexplicable compared to their previous ratings. This mismatch motivated us to conduct a follow-up study (N=20) to confirm these failures in a controlled setting. We analyze interaction data and user feedback through the theories of human judgment to explain the observed failures of the human-AI loop with two key insights: 1) optimization using preference lacks mechanisms to deal with inconsistent and contradictory human judgments; 2) machine outcomes, in turn, influence future user choices via heuristic biases and loss aversion. In this light, we conclude the paper by proposing descriptive UI design guidelines for mitigating these practical problems.

2 RELATED WORK

We will first briefly summarize recent work in BO on preference learning and polygon reduction methods in geometry processing, which drove our system implementation. Then, we will discuss the fact that recent HCI research emphasized the benefits of using HITL to support user or algorithm efficiency, but that, on the other hand, little is known about what might keep such a system from working properly.

2.1 Learning Human Preferences using Bayesian Optimization

As processing human subjective input is challenging, modeling human choices [10, 16] is a research topic in preference learning. To exploit preferences, formal models consider a human rater as an unknown utility function that produces ratings, such as on a 5-star scale, concerning evaluation instances.

The optimization of control parameters based on feedback from a human relying on domain knowledge can be described as an optimization problem: Search an optimal *parameter set* $p^* \in \mathcal{P}$ such that $p^* = \text{argmax}_{p \in \mathcal{P}} h(M(p))$ where p is any parameter set in the *parameter space* $\mathcal{P}$, $M(p)$ is the instance generated using parameter set p, and the *preferential function* h returns the human rating of system output $M(p)$.

When querying the function h is costly, for example, because it involves asking a human for a preferential decision, BO is often selected [5, 6, 39, 44] and has provided a wide range of successful applications [21, 39, 54, 65]. BO actively learns a *posterior* from human preference based on the collected ratings and can then predict an estimated optimal p^* in the next iteration step. Thus, this method captures how artists usually develop their expertise with a new system by exploration and exploitation, resulting in an overall improvement after some iterations.

As a variant, preferential Bayesian optimization (PBO) [20, 39, 44] is an improved version of BO. It evaluates preference on paired samples with an adaptive reference point. The reason for using PBO is that humans are better at differentiating paired samples than they are at determining absolute values [6, 7, 10] according to Thurstone's law of comparative judgment [57]. Therefore, it is considered better for use in HITL systems, and we integrate PBO to learn human preferences.

2.2 Polygon Reduction in Geometry Processing

For the specific geometry processing problem, polygon reduction, presented in this paper, we briefly discuss its recent advances and the relevant methods we utilized to build our system. The geometry processing "No-Free-Lunch" theorem [63] states that not all geometric properties can be well preserved simultaneously in discrete instantiations of a smooth geometry. Therefore, different processing tasks have specialized algorithms and corresponding configurations, such as soft or hard geometries. In general, polygon reduction methods can be categorized as *local decimation*, *global remeshing*, or a weighted combination of both.

Local decimation means that neighbor vertices and edges are greedily removed. These methods date back to the last century [17, 18] and have also been used for levels of detail (LOD) generation [22], even baked into hardware rendering pipelines [46]. They are efficient but contain ill-posed cases with results depending on the implementation. Instead, the general idea of global remeshing [2, 14, 36, 48] is to define a *directional field* as constraint boundary conditions on a *Poisson equation*, and then minimize an artificial energy function. After minimization, a new target mesh can be reconstructed from scratch using the solved solution. Computationally, this is much more costly and cumbersome, but the resulting mesh quality is much better than that from local decimation. State-of-the-art practical solutions, such as Karis et al. [33], use a weighted combination of both that balances the processing speed against quality. A large mesh can be split into smaller ones, then processed using mixed global [23, 25] and local [18, 22] methods, but this also introduces the new problem of cutting a mesh. We refer to Metis [34, 49] as a mature solution for graph partitioning.

2.3 Decision Error in Human Judgments

Economics widely studied decision-making when choosing a preferred item from several alternatives. The expected choice utility maximization [43] forms the theoretical basis. It describes a standard economic model on a finite number of decisions but assumes that individual beings behave rationally. In psychology, Simon [56] proposes the concept of *bounded rationality* and proposes to replace this assumption, as rationality is only limited, and decisions are made by *satisficing*. Later, *prospect theory* [30, 59] empirically demonstrated human judgments in reality when trying to maximize a certain utility function (wealth) in risky situations (e.g., under time pressure) and explained the behavior with bounded rationality. The *heuristic biases* constitute a key source of general decision error. Tversky and Kahneman [58] showed that in any decision under uncertainty, System 1 (fast and instinctive thinking) tends to override System 2 (slow and rational reasoning) [26], hence creating a statistical bias on the decision. More specifically, 1) *representativeness* substitutes the most readily accessible examples to form a decision, 2) *availability* uses mental shortcuts, and 3) *anchoring* as a conclusion bias describes that initial information has a consequence on a later decision.

In addition to heuristics, other effects can also influence judgments: 1) in a utility maximization context, *diminishing returns* [55] may occur as wealth increases and marginal utility decreases; 2) *loss aversion* [29], as part of the *endowment effect* [27], describes that people prefer to retain an owned property rather than to acquire an alternative, potentially better one. People hence tend to stick to seemingly safe decisions when a potential gain would require more risk. 3) In the present understanding, combined with a statistical view [4, 19], systematic noise descriptively shapes another form of decision error that contributes equally to judgment error as individual bias [28]. The decision noise components [28] break down into *level noise* (decision variability between groups), *stable pattern noise* (contextual bias within groups), and *transient noise* (purely occasional).

2.4 Human-in-the-Loop Systems

The HITL strategy may be applied in different contexts, which connect to different fields, including personalization, co-creation, and decision-making support.

From the human computation [50, 60] perspective, a HITL system may be designed to use crowds [24] as human processors to solve system tasks that neither machine nor human can solve independently. Although using collective intelligence has been largely verified to be beneficial for crowdsourcing tasks [31], there are several identified challenges [53] to integrating human computation, which highlighted challenges such as user motivation, sustainability, and input bias. To motivate users to contribute, researchers have used a Game-With-A-Purpose (GWAP) approach [40], but turning a task into a game could be another challenging design problem. Dealing with diverging opinions within small crowds may be difficult because tasks might require a certain level of expertise. Moreover, HITL systems using crowds may suffer from malicious inputs [47] and lead the entire system toward using biased inputs when the initial samples lack trust. Particularly for design-related tasks, crowd opinions may not fit individual interests and needs regardless of data bias. Hence using crowd-powered design systems [38] is considered limited when individual customization has a higher priority.

In a personalized context, Buschek et al. [8] examined the potential pitfalls for achieving user interests in the co-creation context. The limitations on the machine side, identified as lack of machine *creativity* [41], and *usability* [37], and thus, highlight a biased AI with trained system bias but lacks discussions on the source of bias and the mismatch between individual expectations and system abilities. In terms of mismatched expectations, Eiband et al. [15] reported that users might intentionally provide flawed inputs when a system fails to achieve their satisfaction in everyday intelligent applications. As a follow-up, however, Völkel et al. [61] showed that a user must exhaustively provide noisy feedback to confuse an intelligent system. Still, they lack verification and interpretation as to whether the repeated unsatisfactory results come from system limitations or human behavior change. For AI-assisted decision-making scenarios, *trustworthiness* becomes a primary social concern regarding *reliability* in areas where a decision is vital, such as clinical decisions [9, 35]. Still, it is *implicitly assumed* that the human involved eventually makes a rational decision over *subjectively* untrusted AI outcomes. Factors such as algorithm aversion [12] were confirmed to indicate that users are more biased [51] towards human results and produce considerable noise even in the judicial area [13].

Although prior research [11, 21, 38, 42, 64, 66] that involves HITL strategies have shown human knowledge to be helpful for a machine to learn, previous literature rarely discusses the circumstances under which HITL could shine. Especially when users intentionally or unintentionally provide defective or uncertain inputs, it is unclear whether the system can continue to process it effectively and whether other cascading effects will be triggered. Our paper addresses this research gap and shows challenges that can arise when exploiting individual preferences in a HITL system in practice.

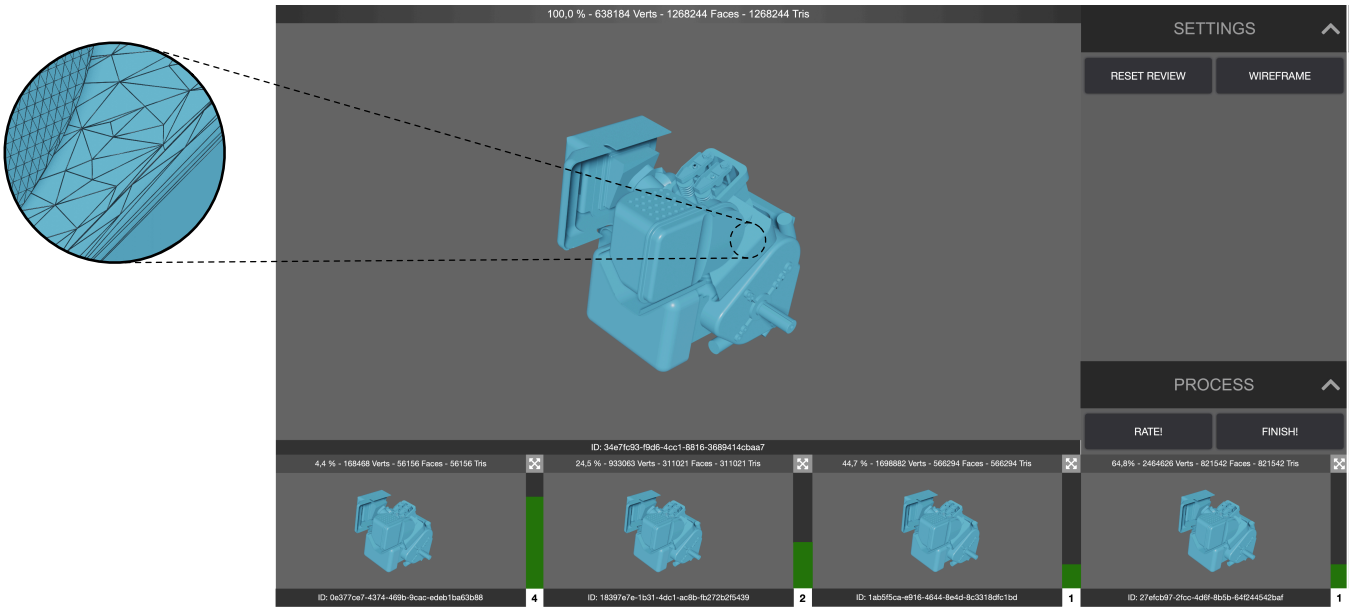


Figure 2: The UI for users to rate variant models. An artist can move each mesh variant to the bigger, central view and activate a wireframe for detailed quality inspection, which was considered necessary in previous work [3] because of the alignment of mesh edges is considered a quality metric. Artists must make a professional choice between visual rendering quality and wireframe quality, especially in cases that may sacrifice a bit of wireframe quality for substantial gains in polygon reduction (lower number of triangles).

3 SYSTEM DESIGN AND USER WORKFLOW

We approached the polygon reduction problem in a joint project with an industrial collaboration partner and were motivated by the goal to optimize their artists' daily tedious manual tuning workflow. One of the benefits of having simplified 3D mesh models is to reduce rendering complexity due to fewer primitives required in a render pass. 3D mesh-based models in the project ranged from a single water-tight mesh layer to models that were either manually sculpted or automatically converted from solid geometry data formats, containing hundreds of sub-meshes with millions of polygons in total. Domain experts processing these mesh models usually require days or weeks of extensive manual work.

We expose the core functionalities of our system as Web APIs that run the polygon reduction on a GPU server. The developed frontend UI by our industrial partner uses Unity[1] as shown in Figure 2. We engineered a hybrid local/global algorithm based on state-of-the-art research [18, 25] controlled by nine different parameters for the polygon reduction itself. We determined the initial parameters by domain heuristics in a few pilot tests with our industrial partner.

In our designed workflow, an artist can first upload an original model. The server then simplifies the uploaded model under different parameter settings in the background. When it has computed all alternatives, taking between seconds and minutes, they are downloaded back into the UI. When the artist indicates their ratings of model quality, the system learns from these judgments and continues the process again to generate more optimized models. The rating scale for judgments is 0 (*skip*, meaning not considered due to faulty geometry), 1 (*terrible*) to 5 (*excellent*). Without loss of generality, if the human decided ratings for the four variant models $M_i (i = 1, 2, 3, 4)$ are 3, 4, 5, and 1, then this represents six preferential choice relations: $M_1 \preccurlyeq M_2$, $M_1 \preccurlyeq M_3$, $M_2 \preccurlyeq M_3$, $M_4 \preccurlyeq M_1$, $M_4 \preccurlyeq M_2$, and $M_4 \preccurlyeq M_3$ where $\preccurlyeq$ means "is less preferred." 3D models in the next iteration are optimized based on these relations using PBO which made us expect [20, 44] the system to converge to the desired outcome quickly. Note that the model quality is *not* equivalent to just visual rendering quality but also involves edge flow and other geometric properties [3] that require a subjective decision. Hence human judgments in this task involve visual rendering quality, mesh wireframe alignment quality, and other technical metrics such as volume-preserving, which is also why this task is more preferred to query human opinions than done by full automation.

[1] https://docs.unity3d.com/2020.3/Documentation/Manual/UIToolkits.html

4 USER STUDIES

To evaluate the effectiveness of our system, we conducted two studies; first, a field study with our industrial partner, and second a lab study to verify the effects of the field study. In the field study, we used our system in a real-world setup with our industrial partner, who aimed to optimize their process in customer projects. Here, we had the opportunity to run a field study where designers used the system in their daily workflow. However, due to the uncontrolled environment of field study, making in-depth assessments and conclusions on specific aspects can be hard. Thus, we additionally ran a lab study to understand the failure cases of our system to verify our findings further under controlled conditions.

4.1 Field Study

For the field study, we first conducted a set of small pilot experiments to fine-tune our system's parameters to fit the partner's needs and their customer projects.

Participants. We recruited two full-time 3D technical artists from our industry partner to gain insights into the newly developed workflow involving our interface. Both are male, aged 25 and 35, one has more than three years of experience, and the other has more than eight years of experience in the 3D industry.

Procedure. The two experts used our system almost daily to evaluate model quality during polygon reduction. However, they were not restricted to our interface and could also use further software aids (as in their previous workflow) for the model quality inspection, e.g., for accessing more professional curvature visualizations. When using our tool, the loaded 3D model was computed into four variants. After the experts finished their evaluation (either inside our interface or externally), they rated the models in our interface. These ratings were used for the next optimization iteration to generate new variants (see Figure 1). The rating process terminated when the experts found the results satisfactory or reset it.

Collected Dataset. During the three months of the study, we collected 549 evaluation sequences as a field study dataset. This corresponds to 4.5 evaluations per expert and workday. Of these, 415 sequences terminated in the first iteration without any preference optimization requested. The remaining 134 sequences (number of iterations: $\mu = 4.1, \sigma = 4.2$, range 1-23) contain sequential preference ratings. The collected 3D models included various mesh types, including organic, soft surfaces, hard technical surfaces with sharper angles, and combinations, such as machinery parts (see Figure 2) with both smooth, flowing lines and hard, mechanical edges.

4.2 Lab Study

As a lab study, we conducted a within-subjects user study to further understand our system's use with specific assistance information and behaviors in a larger user group with different backgrounds in 3D.

Participants. We recruited 20 participants using convenience sampling (7 female and 13 male; age $\mu = 27.0, \sigma = 8.8$, range 18-62). Among them, four had more than one year of industrial experience in 3D modeling, and all others had no experience.

Figure 3: Models that are selected in the lab study, from left to right: *monkey, teapot, rose, cow, pumpkin*. Each model was rendered *with* and *without* wireframes separately.

Procedure. We first welcomed participants and explained the study, answering all open questions before they signed the consent form. Then, participants were presented with different 3D objects in every evaluation session. The overall procedure in terms of rating and termination process in each evaluation session was similar to in the field study. In detail, we asked participants to balance the trade-off between polygon reduction and quality loss. Thus, they had to indicate their preference using ratings to optimize models iteratively. Participants could terminate fast in a few iterations if satisfied or were stopped at the 11th iteration to limit the time of participation. After each evaluation session, we asked participants 7 questions (see Section 8) to indicate their satisfaction and provide overall subjective feedback. We selected five different 3D models, as shown in Figure 3, and each model was rendered *with* and *without* wireframes (instead of allowing users to activate them freely). We picked these five models with the two wireframe representations to ensure our results generalize beyond this small set of objects. We displayed the order of these 3D models and their wireframe display using a Latin square design to avoid learning and fatigue effects. Therefore, we collected $5 \times 2 = 10$ evaluation sessions in total for each participant. On average, each participant spent 90 minutes in the entire study.

Collected Dataset. We collected 200 evaluation sequences (number of iterations: $\mu = 5.1, \sigma = 2.9$, range 1-11) by design, and all sequences involved at least one preference optimization. The selected model covers a similar spectrum of models as the experts had experienced in our field study. These models were also simpler than complex real-world models to reduce the time of machine optimization and participation waiting time.

5 RESULTS

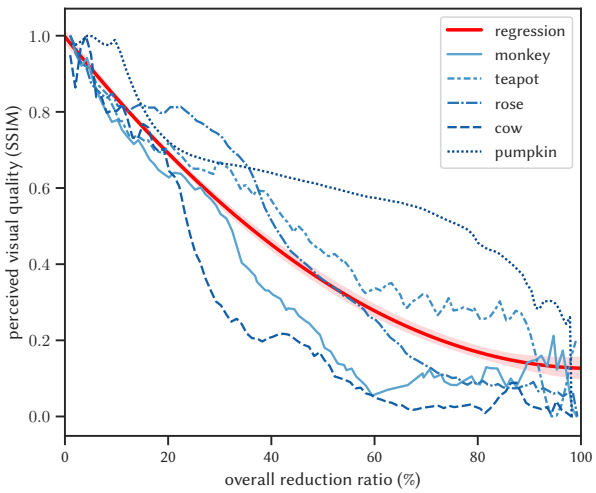
Below, we report our analysis of the participants' rating process using collected data from the two studies, also in comparison to each other, and show that if participants are not rated by pure random, they at least behave highly unstable and inconsistent in the rating process. Then, we show selected example cases from the collected data that were also discussed with experts in hindsight concerning why they made a particular rating choice.

5.1 Human-AI Mutual Interventions

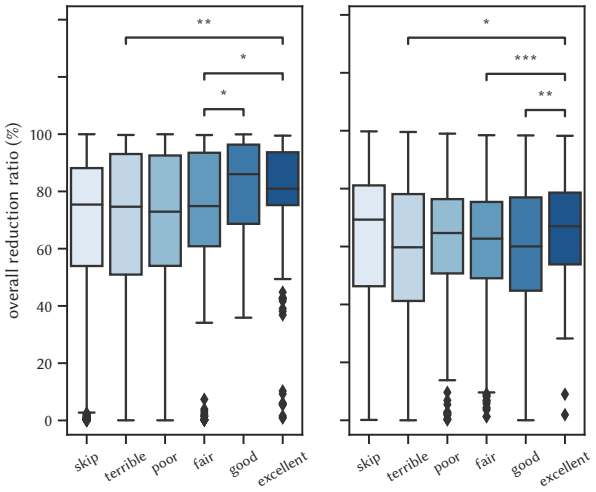
Overview. In the *field study*, from the 134 sequences with preferential ratings, only 16 sequences (11.9% = 16/134) produced a satisfactory outcome. In the *lab study*, among the collected 200 sequences, only 97 sequences (48.5%) were terminated with a satisfactory outcome. Both studies suggest a high failure rate in optimizing HITL outcomes.

Effectiveness of Human Judgments. Figure 4a shows a second-order polynomial regression between a human perceived quality of 3D models used in our lab study and the overall reduced amount of polygons. We assess perceived visual quality using an average of multiple *structure simularity* (SSIM) [62] that compares the rendered visual quality between the reduced and original 3D model in five different camera views. The *reduction ratio* represents the removed polygon count of a resulting model divided by the total polygon count in the original model. Figure 4b shows the rating distributions in the two studies regarding reduction ratio.

We used Kendall's τ coefficient to measure the ordinal association between the reduction ratio and rating scale. The result shows



(a) The relation between the reduction ratio of models in lab study and the human perceived visual quality (using SSIM, normalized), and the red solid smooth curve shows a second-order polynomial regression.



(b) Field (left) and lab (right) studies' preferential ratings against overall reduction ratio (higher means stronger reduction), * ($p < .05$), ** ($p < .01$), * ($p < .001$).**

Figure 4: Overview of the visual influences of polygon reduction and collected rating data.

a significant correlation ($\tau = 0.07, p < .001$) in the field study whereas no significance ($\tau = 0.004, p = 0.71$) in the lab study. This suggests that field study experts tend to give higher ratings to highly reduced models, but lab study is more diverse. For a more fine-grained measure between rating scales, we also used Mann–Whitney U tests to check for dependencies between different rating scales and the reduction ratio: 1) We found an significant difference between *fair* ($Mdn = 74.88$) and *excellent* ($Mdn = 80.95$) ratings ($U = 8756.0$, $p < .001$) and a significant difference between *terrible* ($Mdn = 74.65$) and *excellent* ($Mdn = 80.95$) ratings ($U = 10219.0$, $p = .003$), i.e. in cases where reduction ratio was positively correlated with rating. On the other hand, we found no significant differences in reduction ratio between *terrible* ($Mdn = 74.65$) and *poor* ($Mdn = 72.93$) ratings ($U = 19705.0$, $p = .62$) or *good* ($Mdn = 80.03$) and *excellent* ($Mdn = 80.95$) ratings ($U = 3602.0$, $p = .37$), i.e., where there would have been a negative correlation. *In sum, this suggests that the collected ratings are effective to the highly reduced models, and the reduction ratio is one of the effectively relevant factors in human judgments.* 2) We found no significant differences in highly reduced models between good and excellent in the field study, but a significant difference between *good* ($Mdn = 60.03$) and *excellent* ($Mdn = 66.99$) ratings ($U = 131031.0, p = .002$) in the lab study. Although the field study had fewer users, this could also be interpreted such that *experts in the wild use other quality metrics, which lab participants with less expertise overlook.* 3) Models rated as *good* and *excellent* have higher mean reduction ratio in the field study ($M_{\text{good}} = 81.68, M_{\text{excellent}} = 75.52$) than in the lab ($M_{\text{good}} = 61.10, M_{\text{excellent}} = 64.84$, also see Figure 4b), which suggests that *lab study participants are easier to satisfy by the system outcomes than expert artists.*

Stationarity and Trends of Data. Figure 5 compares how an ideal (far left) and three actual rating distributions (the rest) drift over time: In our context, since the objective of using PBO is to search a polygon reduction configuration to maximize the human ratings [20], ideally, in a successful exploring and exploiting sequence of preferences, and the mean rating score should *increase* and drift from low values with high variance towards higher values with lower variance (*non-stationary* and with an increasing *linear trend component*). However, the actual sequence shown (as most others) stagnates and fluctuates back and forth. From the 200 sequences collected in the lab, 79 continued to at least four iterations (required for the subsequent trend test), and we tested them using an Augmented Dickey-Fuller test. Results show that 36 rating sequences are stationary ($p < .05$). In the remaining 43 non-stationary sequences, a Mann-Kendall test found only four significant increasing trends in the mean rating score ($p < .001$) and only one significant decreasing trend of rating variance ($p < .05$). Another Mann-Kendall test found that only three sequences had increasing and six sequences decreasing trends regarding the machine-estimated optimal reduction ratio ($p < .05$).

In summary, all these results imply for the optimized process, that 1) on the human side, the rating behavior does not improve over iterations; 2) on the machine side, the optimized reduction ratio using preferential choices does not improve over iterations. This suggests that *the human-machine loop as a whole is kept from terminating and fails.*

5.2 Semi-structured Interviews with Experts

In the semi-structured interviews, we discussed with the two expert artists, case by case, inconsistently judged models and why they made a certain (contradicting) choice. Figure 6 shows three of the discussed models. Figure 6a contains a head model and a more straightforward example of a CAD-converted cylinder that is cut by a sphere. Below, we discuss three example cases in more

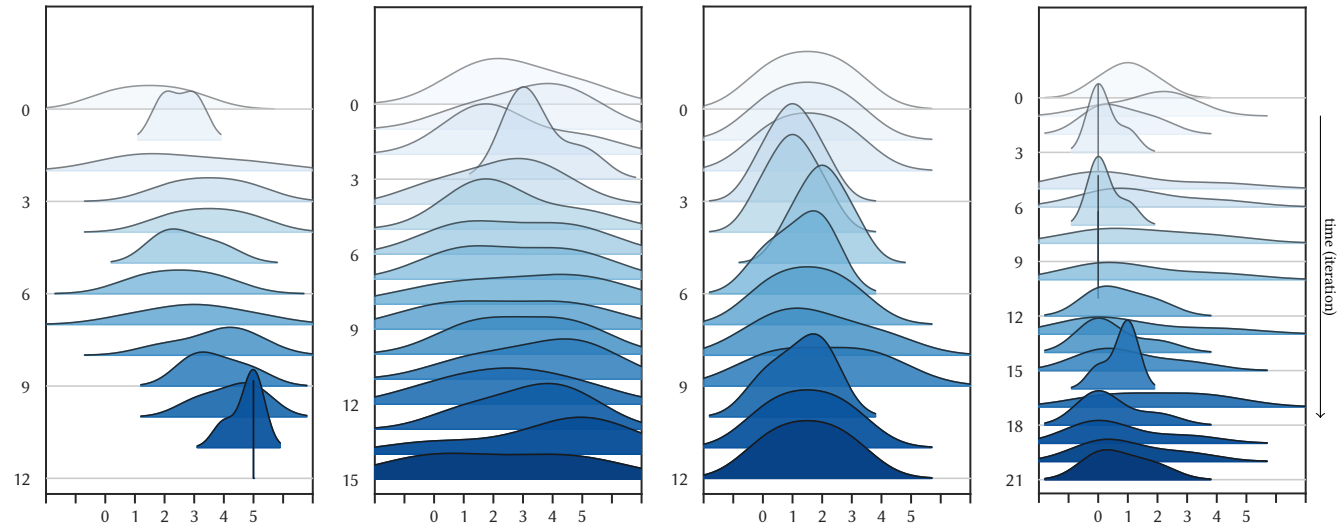
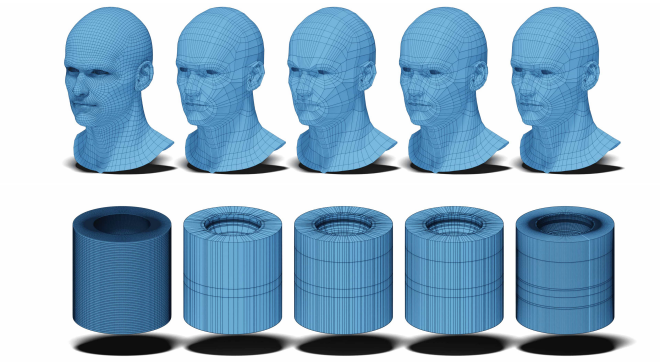
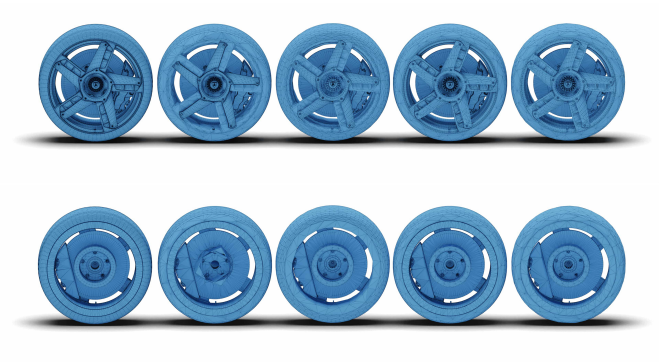


Figure 5: Comparing an ideal (far left) and actual (the others) rating distributions over time (from top to bottom), the bottom axis 0 to 5 represents the rating scale. An expected rating distribution should move to the right side over time if users are more satisfied with the results, but the actual preferences drift back and forth between 0 (*skip*) and 5 (*excellent*).



(a) In the same iteration, original (far left) and 4 processed models: 1) Head model: the last 2 models are identical, ratings (left to right): 5, 3, 4, 1; 2) Cylinder model: the middle two are almost identical. Ratings: 5, 4, 3, 2.



(b) Original (far left) and 4 reduction variants, all rated as 5 in the same iteration. The bottom row shows the backside of the models. Objectively, the middle left wheel contains faulty geometry and should receive a 0 (skip) rather than a 5 (excellent).

Figure 6: Selected examples that were discussed in the semi-structured interviews.

detail: Ironically, the two head models on the right are identical but received an entirely different rating scores of 4 and 1 from the same rater in the same round (*case 1*). The middle right cylinder (rated 3) compared to the middle (rated 4) contains fewer polygons and better symmetries but received a lower rating (*case 2*). As a slightly more complex case (Figure 6b), a wheel model with reduced variants that all were rated a 5 in the same iteration, but the middle left one did reduce to faulty geometry on the backside (*case 3*), which the artist had missed.

In case 1, *artist A* confirmed that his ratings strongly depended on what he had seen before and admitted that he tended to give one model a terrible rating in each iteration due to previous rating experience. He also mentioned that "*...I sometimes stopped giving a higher score because I had decided on a different objective*" in processing the model (e.g., to go for more visual quality but less reduction). In case 2, *artist B* argued that he had scored the middle cylinder mesh higher than the middle right one because "*..the usually difficult inner hard edges were handled better.*." in that case. However, the middle right model has an objectively higher reduction ratio, and both contained similar defects on the inner hard circular edge. Their differences are only at a technical level. Furthermore, he explained that in case 3, he did not notice the flaw at first sight and rated the wheel a 5 simply because it was shown from the front. He had made a quick decision based on the visible mesh quality of the tire and based on a similar experience, which is an example of a reasonably simple oversight with potentially harmful consequences. The artist

also explained that after many iterations, "*...it gets frustrating to see the more flawed output after I already had seen a partially good result.*"

6 DISCUSSION

Although our evaluation does not examine any entangled causality but only statistical correlation, it is likely that the observed system failure initially starts from human error as the system was initialized with the same prior in each of the sequential evaluations (using a Matérn kernel ($\rho = 2, \nu = 2.5$) with the statistical properties of being isotropic and stationary [52, 54]). Since errors are further propagated and amplified to system outcomes, we combine theories regarding human decision errors to reflect on and explain our findings.

6.1 Pitfalls

Error sources on the human side. Based on our observed instability and expert feedback, we argue that the human cognitive errors, which either occur internally or are influenced by the system outcomes, are a crucial part of the overall system uncertainty:

1) *Heuristic biases.* a) The *anchoring bias* explains that earlier experience influences human decisions, including earlier system output and other context factors, such as background knowledge or expertise. In case 1 (see Section 5.2), the artist confirmed that his evaluation depended on meshes he had seen before. b) The *availability bias* explains that judgments are based on the quickly accessed memories of relevant examples. The case 3 matches this bias as the artist decides based on his professional experience. c) *Representativeness* shows that decisions made by substitution examples may be occasionally biased. The case 2 shows this behavior because the actual decision used mental shortcut and was made by judging another similar case.
2) *Loss aversion and endowment effect.* a) Users may become more critical after observing several good results from an intelligent system. Users might stick to what they know and are familiar with and reject newly proposed and objectively good choices, which leads to more negative ratings later in the process. This may explain (case 3) why artists stuck to mediocre choices in intermediate stages instead of moving to a broader (but more risky) range of variations. b) The software functionality (in our case, this is the software pre-configured camera angle for displaying meshes) as a task context may override information and influence the validity of the human judgment. This also explains the unexpected rating of the wheel model in case 3. c) Human preferences change over time and may become inconsistent when interacting. A present rating choice also carries long-term influences, in contrast to being just local. In our case, this is explained by the anchoring bias. It was expected to be addressed in PBO, which uses comparative judgments, but as the three discussed cases show, artists still keep previous experience (either accumulated expertise or short-term outcomes) in mind, which changes their preferential choices.
3) *Diminishing returns.* Judgments may lose precision and contain increasing noise after humans have seen increasingly or partially good results. Hence, preference exploitation may become less effective, and the HITL system can no longer benefit from human knowledge. In our case, when artists had seen a certain number of increasingly better meshes, they were less sensitive (case 3) to further improvements by the algorithm. In contrast, they even gave more critical scores for the occasional poor results.

Error sources on the machine side. The other part of overall system uncertainty comes from the underlying algorithm and is emphasized by user errors:

1) *Stable preference assumption.* The system performance in a HITL system suffers from the model assumption, and the outcomes may be undesired due to an invalid optimization. We observed that human judgments produce strongly local, partially global, time- and context-dependent errors, even with permanent goal changes (case 1). This violates the prerequisites of any optimization technique that assumes a unique and stable utility function, including PBO. More importantly, human judgment is a fragile function to optimize for, and the commonly used *independent and identically distributed* (i.i.d.) assumption in these algorithms does not hold in reality for humans. In turn, we need to generally rethink basic assumptions and approaches in the design of HITL systems. We should be more explicit about under what circumstances they can be applied appropriately to detect and exploit changes in latent user preference distributions and systematic errors.
2) *Complete preference assumption.* The underlying optimization still implicitly assumes a user always has a complete preference, meaning that users are deemed to be able to provide a rating to reflect their preference consistently. In the current design, users rate four models instead of requiring them to choose one of the best. This design can mitigate the completeness assumption violation, as selecting the best might not be possible if comparing objects are not entirely comparable and involves multiple optimizing objectives. However, as the optimization process continues, human raters may lose their preference for rating different models due to bounded rationality.

6.2 Potential Countermeasures

The heuristics are rather hard to detect by the machine since human ratings may not be entirely judged for consistency (otherwise, the machine could provide ratings on its own, entirely defying the idea of HITL systems). Nevertheless, we propose several design guidelines to at least mitigate different types of decision noise as discussed in Section 2.3 thereby may be more parametrically guiding users in further optimization steps:

1) Reduce *level noise*: Provide a timeline to include intermediate results saved by users and allow them to return to those earlier results for comparison. This could help the user to compare new results to known ones and support a more objective comparison across iterations. It could also reduce user frustration and fear of losing the achieved quality, thereby mitigating problems from loss aversion and violated system assumptions;
2) Reduce *stable pattern noise*: Indicate the optimization intention to the user, such as current system steps regarding exploitation and exploration. This could better frame the current context, therefore, mitigate representativeness and availability bias by keeping users from judging based on earlier examples.

3) Reduce *transient noise*: One approach could be to occasionally present results from earlier iterations and check for consistency, although this would also assume stable preferences and require more user iterations. Another approach could provide more assistive visualization by highlighting the mesh difference between iterations. This could further reduce user workload and help mitigate simple oversight and obvious mistakes when distracted by unchanged parts or overlooking changed parts.

6.3 Limitations

Our UI was consciously simplified to a minimum in order not to distract from judgment and in an attempt to avoid usage complexity and improve overall usability. In the field study, due to the limited number of users and to not further confuse users by silently changing system behavior, we did not run any forms of A/B testing. Although the subjects could explore the quality of the entire mesh by features such as enabling the wireframe, this might still have been too restrictive and lacked information about the changes. Highlighting the crucial changes may be helpful, but it also lacks the ability to customize references to show the difference between different proposals in a sequential optimized workflow. In hindsight, we learned that it might be useful to let users specify which parts of the mesh led to a particular rating. Next, we wanted to ensure the generalizability of our results and thus, selected five models with two wireframe representations. On the other hand, our results may still suffer from a selection bias in the models we used. Lastly, conducting a simulated user study [32] and designing further statistically verifying the decision biases might also be helpful to compare simulated human inputs with controlled noise and the actual decision behaviors.

7 SUMMARY AND FUTURE WORK

In this work, we discussed a HITL system where an optimization process in the background exploits sequential user choices to optimize the system's future outcomes iteratively. Our case study provides evidence of challenges to human-AI loops in practice, produced by mutual negative influences. Based on interaction data in the field and lab and discussions with expert artists, we reflected on concrete influences that can break preference-optimized HITL systems, namely by 1) human decision biases and noise, 2) system capabilities to deal with them, and 3) subsequent impact on future human inputs.

The findings provide some answers to our initial research questions: RQ1) The constraints of cognitive effects and the underlying algorithm, such as *heuristic biases*, *endowment effect*, *diminishing return*, and *violated system assumptions*, can be used to explain our empirical observations. Supporting polygon reduction tasks using the HITL strategy requires resolving these issues. RQ2) The observed constraints also apply in a similar HITL context, and we proposed descriptive UI design directions as promising countermeasures to prevent HITL system outcomes from being highly unstable and eventually non-satisfactory.

For future work, we expect to verify the proposed countermeasures in various scenarios and test for similar or different phenomena in other domains. We will also perform a systematic analysis of the basic building blocks that include other HITL systems and related cognitive factors as a foundation to inform guidelines for more error-tolerant HITL systems.

8 OPEN SCIENCE

We encourage readers to reproduce and extend our results. Our system tools, technical specifications, anonymized datasets (without proprietary models and associated user ratings), and evaluation scripts are open-sourced that may be found in https://changkun.de/s/infloop.

ACKNOWLEDGMENTS

The authors would like to thank our industrial partners Stefan Sigl, Marco Petrassi, and Marvin Juschus for sharing challenges in their daily 3D workflow; our colleague Kai Holländer for helpful discussions and feedback; to Feng Chen for executing the lab study, and to Prof. Eyke Hüllermeier and Karlson Pfannschmidt for useful discussions in preference learning. 3D mesh artifacts are provided courtesy of WAY Digital Solutions, Jeff H, Jose Olmedo, kenik, yarulesemel, and Stephan Thieme. This work was supported by the Bavarian IuK Program (IUK1805-0004//IUK577/002). Work on this project is also partly funded by the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation (bidt).

REFERENCES

[1] 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

[2] David Bommes, Bruno Lévy, Nico Pietroni, Enrico Puppo, Claudio Silva, Marco Tarini, and Denis Zorin. 2013. Quad-Mesh Generation and Processing: A Survey. *Computer Graphics Forum* 32, 6 (Sept. 2013), 51–76. https://doi.org/10.1111/cgf.12014

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

[4] 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

[5] 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

[6] 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

[7] Eric Brochu, Abhijeet Ghosh, and Nando de Freitas. 2007. Preference galleries for material design. In *ACM SIGGRAPH 2007 posters*. ACM Press, New York, NY, USA, 105–es. https://dl.acm.org/doi/10.1145/1280720.1280834

[8] Daniel Buschek, Lukas Mecke, Florian Lehmann, and Hai Dang. 2021. Nine Potential Pitfalls when Designing Human-AI Co-Creative Systems. arXiv:2104.00358 [cs.HC] http://arxiv.org/abs/2104.00358

[9] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[10] Wei Chu and Zoubin Ghahramani. 2005. Preference learning with Gaussian processes. In *Proceedings of the 22nd international conference on Machine learning (ICML '05)*. Association for Computing Machinery, New York, NY, USA, 137–144. https://doi.org/10.1145/1102351.1102369

[11] Fabio Colella, Pedram Daee, Jussi Jokinen, Antti Oulasvirta, and Samuel Kaski. 2020. Human Strategic Steering Improves Performance of Interactive Optimization. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. Association for Computing Machinery, New York, NY, USA, 293–297. https://doi.org/10.1145/3340631.3394883

[12] 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

[13] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (Jan 2018), 1–5. https://doi.org/10.1126/sciadv.aao5580

[14] Hans-Christian Ebke, Marcel Campen, David Bommes, and Leif Kobbelt. 2014. Level-of-detail quad meshing. *ACM Transactions on Graphics* 33, 6 (Nov. 2014), 184:1–184:11. https://doi.org/10.1145/2661229.2661240

[15] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When people and algorithms meet: user-reported problems in intelligent everyday applications. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 96–106. https://doi.org/10.1145/3301275.3302262

[16] Johannes Fürnkranz and Eyke Hüllermeier. 2003. Pairwise Preference Learning and Ranking. In *Machine Learning: ECML 2003 (Lecture Notes in Computer Science)*, Nada Lavrač, Dragan Gamberger, Hendrik Blockeel, and Ljupčo Todorovski (Eds.). Springer, Cavtat-Dubrovnik, Croatia, 145–156. https://doi.org/10.1007/978-3-540-39857-8_15

[17] Michael Garland and Paul S. Heckbert. 1997. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., USA, 209–216. https://doi.org/10.1145/258734.258849

[18] 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

[19] 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

[20] 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

[21] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M. Pintea, and Vasile Palade. 2016. Towards interactive Machine Learning (iML): Applying Ant Colony Algorithms to Solve the Traveling Salesman Problem with the Human-in-the-Loop Approach. In *Availability, Reliability, and Security in Information Systems*. Springer, Cham, 81–95. https://doi.org/10.1007/978-3-319-45507-5_6

[22] Hugues Hoppe. 1996. Progressive meshes. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM Press, New York, NY, USA, 99–108. https://doi.org/10.1145/237170.237216

[23] Jingwei Huang, Yichao Zhou, Matthias Niessner, Jonathan Richard Shewchuk, and Leonidas J. Guibas. 2018. QuadriFlow: A Scalable and Robust Method for Quadrangulation. *Computer Graphics Forum* 37, 5 (Aug. 2018), 147–160. https://doi.org/10.1111/cgf.13498

[24] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality Management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (Washington DC) *(HCOMP '10)*. Association for Computing Machinery, New York, NY, USA, 64–67. https://doi.org/10.1145/1837885.1837906

[25] 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

[26] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, USA.

[27] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. 1991. Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal of Economic Perspectives* 5, 1 (March 1991), 193–206. https://doi.org/10.1257/jep.5.1.193

[28] Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise*. HarperCollins UK, UK.

[29] Daniel Kahneman and Amos Tversky. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology* 3, 3 (Jul 1972), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

[30] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (March 1979), 263–291. https://doi.org/10.2307/1914185

[31] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining Human and Machine Intelligence in Large-Scale Crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (Valencia, Spain) *(AAMAS '12)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 467–474.

[32] Antti Kangasrääsiö, Kumaripaba Athukorala, Andrew Howes, Jukka Corander, Samuel Kaski, and Antti Oulasvirta. 2017. Inferring Cognitive Models from Data using Approximate Bayesian Computation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1295–1306. https://doi.org/10.1145/3025453.3025576

[33] Brian Karis, Rune Stubbe, and Graham Wihlidal. 2021. A Deep Dive into Nanite Virtualized Geometry. http://advances.realtimerendering.com/s2021/Karis_Nanite_SIGGRAPH_Advances_2021_final.pdf

[34] George Karypis and Vipin Kumar. 1997. METIS: A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices. http://conservancy.umn.edu/handle/11299/215346

[35] Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. 2021. Explainable Artificial Intelligence for Human Decision-Support System in Medical Domain. arXiv:2105.02357 [cs.HC] http://arxiv.org/abs/2105.02357

[36] Felix Knöppel, Keenan Crane, Ulrich Pinkall, and Peter Schröder. 2013. Globally optimal direction fields. *ACM Transactions on Graphics* 32, 4 (July 2013), 59:1–59:10. https://doi.org/10.1145/2461912.2462005

[37] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. *Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[38] Yuki Koyama, Daisuke Sakamoto, and Takeo Igarashi. 2014. Crowd-powered parameter analysis for visual design exploration. In *Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST '14)*. ACM Press, New York, NY, USA, 65–74. https://doi.org/10.1145/2642918.2647386

[39] 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

[40] Markus Krause and Jan Smeddinck. 2011. Human computation games: A survey. In *2011 19th European Signal Processing Conference*. IEEE, Vancouver, BC, Canada, 754–758.

[41] Carolyn Lamb, Daniel G. Brown, and Charles L. A. Clarke. 2018. Evaluating Computational Creativity: An Interdisciplinary Tutorial. *Comput. Surveys* 51, 2 (Feb 2018), 28:1–28:34. https://doi.org/10.1145/3167476

[42] 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

[43] Jacob Marschak. 1974. *Binary-Choice Constraints and Random Utility Indicators (1960)*. Springer Netherlands, Dordrecht, Chapter Economic Information, Decision, and Prediction: Selected Essays: Volume I Part I Economics of Decision, 218–239. https://doi.org/10.1007/978-94-010-9276-0_9

[44] 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://proceedings.mlr.press/v119/mikkola20a.html

[45] Robert M. Monarch. 2021. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI*. Simon and Schuster, USA.

[46] Marc Olano, Bob Kuehne, and Maryann Simmons. 2003. Automatic shader level of detail. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware (HWWS '03)*. Eurographics Association, Goslar, DEU, 7–14. https://doi.org/10.5555/844174.844176

[47] Changkun Ou, Yifei Zhan, and Yaxi Chen. 2019. Identifying Malicious Players in GWAP-based Disaster Monitoring Crowdsourcing System. In *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD) (ICAIBD' 19)*. IEEE, Chengdu, Sichuan, China, 369–378. https://doi.org/10.1109/ICAIBD.2019.8836972

[48] Nico Pietroni, Stefano Nuvoli, Thomas Alderighi, Paolo Cignoni, and Marco Tarini. 2021. Reliable Feature-Line Driven Quad-Remeshing. *ACM Transactions on Graphics* 40, 4, Article 155 (July 2021), 17 pages. https://doi.org/10.1145/3450626.3459941

[49] Federico Ponchio. 2009. *Multiresolution structures for interactive visualization of very large 3D datasets*. Ph. D. Dissertation. Clausthal University of Technology. http://d-nb.info/997062789/34

[50] Alexander J. Quinn and Benjamin B. Bederson. 2011. Human Computation: A Survey and Taxonomy of a Growing Field. 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, 1403–1412. https://doi.org/10.1145/1978942.1979148

[51] Martin Ragot, Nicolas Martin, and Salomé Cojean. 2020. AI-Generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence?. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3334480.3382892

[52] Carl Edward Rasmussen, Christopher K. I. Williams, and Francis Bach. 2004. *Gaussian Processes in Machine Learning*. Springer, Berlin, Heidelberg, 63–71. https://doi.org/10.1007/978-3-540-28650-9_4

[53] Stuart Reeves and Scott Sherwood. 2010. Five Design Challenges for Human Computation. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (Reykjavik, Iceland) *(NordiCHI '10)*. Association

for Computing Machinery, New York, NY, USA, 383–392. https://doi.org/10.1145/1868914.1868959

[54] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. 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

[55] Ronald W. Shephard and Rolf Färe. 1974. The law of diminishing returns. *Zeitschrift für Nationalökonomie* 34, 1 (March 1974), 69–90. https://doi.org/10.1007/BF01289147

[56] Herbert Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* 69, 1 (1955), 99–118. https://doi.org/10.2307/1884852

[57] L. L. Thurstone. 1927. A law of comparative judgment. *Psychological Review* 34, 4 (1927), 273–286. https://doi.org/10.1037/h0070288

[58] 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

[59] Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5, 4 (Oct. 1992), 297–323. https://doi.org/10.1007/BF00122574

[60] Luis von Ahn. 2008. Human Computation. In *2008 IEEE 24th International Conference on Data Engineering*. Springer, Cham, Switzerland, 1–2. https://doi.org/10.1109/ICDE.2008.4497403

[61] Sarah Theres Völkel, Renate Haeuslschmid, Anna Werner, Heinrich Hussmann, and Andreas Butz. 2020. How to Trick AI: Users' Strategies for Protecting Themselves from Automatic Personality Assessment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376877

[62] 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

[63] Max Wardetzky, Saurabh Mathur, Felix Kälberer, and Eitan Grinspun. 2007. Discrete Laplace operators: No free lunch. In *Proceedings of the fifth Eurographics symposium on Geometry processing (SGP '07)*. Eurographics Association, Goslar, DEU, 33–37. https://doi.org/10.2312/SGP/SGP07/033-037

[64] Thomas Weber, Heinrich Hußmann, Zhiwei Han, Stefan Matthes, and Yuanting Liu. 2020. Draw with me: human-in-the-loop for image restoration. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. ACM Press, New York, NY, USA, 243–253. https://doi.org/10.1145/3377325.3377509

[65] 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

[66] Beste F. Yuksel, Soo Jung Kim, Seung Jung Jin, Joshua Junhee Lee, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Ilmi Yoon, Yue-Ting Siu, and Joshua A. Miele. 2020. Increasing Video Accessibility for Visually Impaired Users with Human-in-the-Loop Machine Learning. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382821

[67] 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