THE INTELLIGENCE IN THE LOOP

Empirical Explorations and Reflections

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von

CHANGKUN OU

M.Sc. Human-Computer Interaction

München, den 17.02.2023

Erstgutachter:	Prof. Dr. Andreas Butz
Zweitgutachter:	Prof. Dr. Eyke Hüllermeier
Drittgutachter:	Prof. Dr. Marc Stamminger

Tag der mündlichen Prüfung: 03.04.2023

Abstract

For decades, engineering in computing systems has used a human-in-the-loop servo mechanism. A conscious human being is usually believed, in a rational manner, to operate, assist, and control the machine to achieve desired objectives. Over time, researchers have started to use human-in-the-loop schemes in more abstract tasks, such as iterative interface design problems. However, with the observations and developments in social science, the underlying rationality assumption is strongly challenged, and humans make mistakes. With the recent advances in computer science regarding artificial intelligence, data-driven algorithms could achieve human-level performance in certain aspects, such as audio recognition, image segmentation, and machine translation tasks. The human-in-the-loop mechanism is being reconsidered and reshaped towards an extended vision to assist human decision-making or creativity in the human-computer interaction (HCI) research field.

This thesis explores the boundary for human-in-the-loop optimization systems to succeed and be beneficial. In the interaction loop, machine agents are designed rationally to interact with human beings that may behave using incomplete rational policies iteratively. The thesis first examines and deliberates common principles in mainstream HCI research regarding the advice for building human-in-the-loop systems using existing computation techniques concerning decision-making support, utility-based optimization, and human concepts regarding preferences, satisfaction, and expertise.

To reflect real-world constraints in a human-in-the-loop optimization system, the thesis explores three design problems: text summarization, image color enhancement, and 3D polygon reduction. These design problems are selected to involve human perception and intelligence, aesthetic preference, and rational judgments. Specifically, to understand and analyze the interaction loop, the thesis conducted a series of experiments to study the impact of various building blocks in human-in-the-loop systems that observes exploration and exploitation of human users, including problem context, solution space, reliability of human inputs regarding preference and expertise, and relevant user interfaces for inputs. Combining the findings of the experiments, the thesis revisits vulnerable assumptions that may be largely ignored when designing a modern human-in-the-loop optimization system.

The experiment on the impact of user interfaces narrows down the exploration space of this thesis and empirically demonstrates how different preferential user interfaces influence the overall interaction performance. Based on the findings, subsequent experiments further investigate how human judgments can be a flaw of a human-in-the-loop optimization system. The result shows that, due to cognitive limitations and unrealistic system assumptions, inconsistent and unstable preferences commonly exist in this human-in-the-loop optimization system, resulting in suboptimal machine outcomes and user dissatisfaction, which conflicts with the objective of using a human to gain the expected output.

With a deeper look into human aspects, another experiment attempts to reveal the potential causes, such as involved level of human expertise. The system further tests the usage of individuals with different levels of expertise. Based on the observation and analysis, higher-level expertise leads to lower subjective satisfaction and more interactions, whereas novices terminate faster and also achieve expert-level performance, which not only reveals challenges to utilizing the obtained human insights but also be considered as an indicator to reveal how we can better involve a human in an optimization loop for exploring a solution space.

All these contributions in human-in-the-loop optimization systems lead to a rethinking of the source of intelligence and engage philosophical discussions. These topics eventually approach more fundamental questions regarding the definition of intelligence and how we might succeed in keeping our *intelligence in the loop*.

Zusammenfassung

Seit Jahrzehnten verwendet das Ingenieurwesen in Computersystemen einen “human-in-the-loop” Servomechanismus. Ein bewusster Mensch wird in der Regel auf rationale Weise eingesetzt, um die Maschine zu bedienen, zu unterstützen und zu kontrollieren, um die gewünschten Ziele zu erreichen. Im Laufe der Zeit haben Forscher begonnen, “human-in-the-loop” Schemata in abstrakteren Aufgabenstellungen wie iterativen Schnittstellendesignproblemen einzusetzen. Allerdings wird mit den Beobachtungen und Entwicklungen in den Sozialwissenschaften die zugrunde liegende Rationalitätsannahme stark in Frage gestellt und Menschen machen Fehler. Mit den jüngsten Fortschritten in der Informatik im Bereich der künstlichen Intelligenz könnten datengetriebene Algorithmen in bestimmten Bereichen menschenähnliche Leistungen erbringen, wie zum Beispiel bei der Audioerkennung, Bildsegmentierung und maschinellen Übersetzung. Der “human-in-the-loop” Mechanismus wird im Bereich der Forschung zur Mensch-Computer-Interaktion (MCI) neu überdacht und neu gestaltet, um die menschliche Entscheidungsfindung oder Kreativität zu unterstützen.

Diese Arbeit untersucht die Grenzen für “human-in-the-loop” Optimierungssysteme, um erfolgreich und vorteilhaft zu sein. In der Interaktionsschleife werden Maschinenagenten rational entworfen, um mit menschlichen Wesen zu interagieren, die iterativ möglicherweise mit unvollständigen rationalen Richtlinien handeln. Die Arbeit untersucht und diskutiert zunächst gemeinsame Prinzipien in der Mainstream-Forschung zur Mensch-Computer-Interaktion (MCI) hinsichtlich der Empfehlungen für den Aufbau von “human-in-the-loop” Systemen unter Verwendung vorhandener Berechnungstechniken zur Entscheidungsunterstützung, nutzungsbasierter Optimierung und menschlichen Konzepten bezüglich Vorlieben, Zufriedenheit und Expertise.

Um realitätsnahe Einschränkungen in einem “human-in-the-loop” Optimierungssystem widerzuspiegeln, untersucht die Arbeit drei Designprobleme: Textzusammenfassung, Verbesserung von Bildfarben und Reduzierung von 3D-Polygonen. Diese Designprobleme wurden ausgewählt, um die menschliche Wahrnehmung und Intelligenz, ästhetische Präferenzen und rationale Urteile einzubeziehen. Um die Interaktionsschleife zu verstehen und zu analysieren, führte die Arbeit eine Reihe von Experimenten durch, um die Auswirkungen verschiedener Bausteine in “human-in-the-loop” Systemen zu untersuchen, die die Exploration und Ausnutzung menschlicher Benutzer berücksichtigen, einschließlich des Problemkontexts, des Lösungsraums, der Zuverlässigkeit menschlicher Eingaben bezüglich Vorlieben und Expertise sowie relevanter

Benutzeroberflächen für Eingaben. Durch die Kombination der Ergebnisse der Experimente hinterfragt die Arbeit anfällige Annahmen, die bei der Gestaltung eines modernen “human-in-the-loop” Optimierungssystems weitgehend ignoriert werden können.

Das Experiment zur Auswirkung von Benutzeroberflächen begrenzt den Explorationsspielraum dieser Arbeit und zeigt empirisch, wie unterschiedliche bevorzugte Benutzeroberflächen die Gesamtleistung der Interaktion beeinflussen. Basierend auf den Ergebnissen untersuchen nachfolgende Experimente weiter, wie menschliche Urteile die Schwachstelle eines “human-in-the-loop” Optimierungssystems werden können. Das Ergebnis zeigt, dass aufgrund kognitiver Einschränkungen und unrealistischer Systemannahmen inkonsistente und instabile Präferenzen in diesem “human-in-the-loop” Optimierungssystem häufig vorkommen und zu suboptimalen Maschinenergebnissen und Benutzerunzufriedenheit führen, was dem Ziel widerspricht, einen Menschen zur Erzielung des erwarteten Outputs zu nutzen.

Mit einem tieferen Blick auf menschliche Aspekte versucht ein weiteres Experiment, potenzielle Ursachen aufzudecken, wie zum Beispiel das involvierte Niveau menschlicher Expertise. Das System testet außerdem die Verwendung von Personen mit unterschiedlichen Kenntnisständen. Basierend auf Beobachtungen und Analysen führt höhere Expertise zu geringerer subjektiver Zufriedenheit und mehr Interaktionen, während Anfänger schneller aufgeben und auch eine Expertenleistung erbringen. Dies zeigt nicht nur Herausforderungen bei der Nutzung der gewonnenen menschlichen Erkenntnisse auf, sondern kann auch als Indikator dienen, um aufzuzeigen, wie wir einen Menschen besser in eine Optimierungsschleife einbeziehen können, um einen Lösungsraum zu erkunden.

All diese Beiträge in “human-in-the-loop” Optimierungssystemen führen zu einem Umdenken über die Quelle der Intelligenz und führen zu philosophischen Diskussionen. Diese Themen nähern sich schließlich grundlegenderen Fragen über die Definition von Intelligenz und wie es uns gelingen könnte, unsere *intelligence in the loop* zu halten.

Acknowledgments

Intersubjective experience plays a fundamental role in our constitution of both ourselves as objectively existing subjects.

– Edmund Husserl

It is a memorable journey definitely for me to pursue a Ph.D.

While most of the ideas of this thesis were developed in 2019, the actual writing and publication went through many side tracks and revision cycles, delayed until 2023 by many external factors. Even until the present, the thesis is far away from perfection. As a happy ending, I am delighted to see that some of my thoughts are published, some are under review, but more are still in my mind. All these good old days shaped my current mindset and taught me many lessons regarding communication, presentation, and an epistemology of the world. As acknowledgment, I would like to thank all the people who have helped me in this journey.

Thank you, **Xiuying Shu** and **Shengxin Ou**, for bringing me into the world so that I am able to experience and explore the unknown. I am grateful for your unconditional love, your patience, your understanding, and support over my past 30 years of life.

Thank you, **Yaxi Chen**, for your life-changing support, giving me the chance to study aboard, and earning the chance to pursue a Ph.D. Thank you, **Andreas Butz**, for being a supervisor, for your great support over the past four years, for your advice, feedback, energy, and optimism, and for many insightful and inspiring meetings. You helped me to discover who I am and what I enjoy doing.

Thank you, **Dennis Dietz**, for always being in a great mood, for your kind inviting me in the same office, and created many enjoyable discussions during lunch breaks and evening walks. Thank you, **Carl Oechsner**, for being a particular colleague in the office, organizing so many amazing parties, and many spicy gifts I enjoyed a lot.

Thank you, **Francesco Chiossi**, for too many insightful discussions between us and for too much mid-night passion towards research, you shaped most of my enjoyable time while working on a Ph.D., and interacting with you is always a pleasure, and I learned a lot from you.

Thank you, **Stefan Sigl**, **Marco Petrassi**, and **Marvin Juschus**, for sharing chal-

lenges in your daily 3D workflow, which opened and helped me develop my initial passion for doing research.

Thank you, **Daniel Buschek** and **Sven Mayer**, for your positive energy and enthusiasm, support and feedback, insightful discussions, and great help in improving my communication and presentation skills.

Thank you, **David Englmeier**, for your understanding and help in dealing with the computer graphics teaching, for always being helpful in creating teaching materials, for supportive discussions in problem-solving, and for your great help in improving my communication skills.

Thank you, **Eyke Hüllermeier** and **Karlson Pfannschmidt**, for your kindness in sharing one of the most helpful pieces of feedback I received during my Ph.D. time. Thank you, **Heinrich Hußmann**, for being a great mentor, giving me the opportunities and trust to help lectures in the university, and for many insightful discussions during subway travels between office and classroom. Thank you, **Albrecht Schmidt**, for always being positive, inspiring, and energetic in the office. Thank you, **Yi Xia**, for your cross-country phone calls and discussions regarding life. Thank you, **Yifei Zhan**, for suggestions on the ranking interface implementation and many fun and inspiring dinner discussions, outdoor activities, and trips. Thank you, **Guoliang Xue**, for many fun and interesting discussions regarding software engineering and life. Thank you, **Yong Ma**, for football discussions in the office. Thank you, **Heiko Drewes**, for your insights and discussions regarding expertise. Thank you, **Jingyi Li**, for so many activities and fun times together in travel, conference, climbing, dinner, discussions, etc. Thank you, **Yanhong Li**, for working together on papers and a painting gift created by you. Thank you, **Wen Yang**, for spending a lot of time discussing and building a Go community together, as well as many recording podcasts. Thank you, **Quancheng Rao**, for sharing the energy with me in discussing and building a Go community together, as well as writing and publishing books. Thank you, **Fei Ao**, for sharing many insights regarding network protocols and engineering details regarding building a storage-free system. Thank you, **Robin Welsch**, for discussions of multilevel modeling during CHI time. Thank you, **Hai Dang**, for the discussion of large language models during CHI time. Thank you **Kai Holländer**, for your generous help in providing feedback to improve my publication. Thank you, **Linda Hirsch**, for introducing me to an interesting visualization of research paper classification, and for a lot feedback regarding paper writing. Thank you, **Fiona Draxler**, for the feedback on measuring language proficiency during our CHI preparation meeting. Thank you, **Nađa Terzimehić**, for being the expert of good taste and for the special 2020 Christmas gift. Thank you, **Pascal Knierim** and **Jesse Grootjen**, for a lot of support in providing coffees in the office, without them, we cannot even work. Thank you, **Sylvia Rothe**, for co-supervising an excellent student and contributing many insights in terms of film making. Thank you, **Florian Bemmann**, for sharing tons of insights regarding outdoor activities. Thank you, **Steeven Villa**, for inviting me to a lot of your interesting research projects. Thank you, **Thomas Kosch**, for many inspiring discussions on what HCI research is and for sharing a lot of interesting work you did. Thank you, **Luke Haliburton**, Thank you, **Beat Rossmy**, for sharing many interesting insights regarding music and artistic creation. Thank you, **Gesa Wiegand**, for discussions of future planning on the stairs Thank you, **Annika Kaltenhauser**, **Sarah Prange**, **Lukas Mecke**, and **Tobias Seitz** for the enjoyable experience during IDC and CHI trips. Thank you, **Tony Zhang**, for your insights regarding decision-making support. Thank you, **Thomas Weber**, for many interesting discussions regarding software developments. Thank you, **Christian Mai**, for your insights regarding career planning. Thank you, **Malin Eiband**, for many interesting discussions about intelligence and creativity. Thank you, **Jerry Lin**, for the discussion of pairwise Gaussian processes during CHI time. Thank you, **Clemens Damke**, for discussions of ranking analysis. Thank you, **Jue Li**, for discussions regarding the Chinese research community during dinner. Thank you, **Xiaomei Qu** and **Tao Liu**, for a lot of discussion on the experience of living in Germany and for sharing your past career experience in doing research.

Thank you, **Rainer Fink**, for great help with all things technical, and I will never have the chance to develop experience and success in maintaining legacy systems. Thank you, **Franziska Schwamb** and **Christa Feulner**, for dealing with official documents and for much support in reimbursing travel expenses.

Thank you to our students, **Johannes Merkt**, **Julius Girbig**, **Christian Schmidt**, **Feng Chen**, **Kehong Deng**, **Benjamin Sühling**, **Gerhard van Nooy**, **Nicolas Mogicato**, **Oliver Möller**, **Kevin Nsieyanji**, **Zihan Kong**, **Elena Liebl**, **Shiyi Gou**, **Samuel Eiler**, and others for many great projects and your hard work.

Thank you to whom is not mentioned here for your support and help in this journey. Without any of you, this journey will never be complete.

Collaboration Statement

The thesis is based on research projects that would not have been possible without the excellent support of my supervisor and colleagues. In appreciation, I consciously avoid using “I” throughout the main body of this thesis to avoid subjectivity. Here, I clarify my contribution in favor of the formal clarity of this “disclaimer.”

Contribution of colleagues: The following statements relate to those publications in which the respective person appears as a co-author. The research was conducted closely with my supervisor Andreas Butz, who provided initial feedback and contributed to the final edits of the work included in the thesis. My colleague Sven Mayer provided feedback on improving the clarity of the writing, study design, research communications, and manuscripts. My student-era supervisor, Daniel Buschek, motivated me for one more submission attempt and contributed to the discussions during implementations, which helped towards the eventual publication.

Content and presentation: I developed all core ideas and concepts presented in this thesis, supported by feedback, discussions, and suggestions from my supervisors and colleagues. I implemented all analysis scripts, prototypes, and software. I wrote the complete text for all publications, computed all data analyses which appear therein, and created all figures and tables. My co-authors contributed feedback and language-level edits on the manuscripts. I personally revised all content, integrating their feedback. I am responsible for the thesis’s overall structure and the presentation of the results.

Table 1 clarifies contributions of others to individual projects and publications.

Publication	**Contributions of Others**
Ou et al. [113]	S. Mayer contributed to the study design and the clarification of research hypothesis. D. Buschek contributed to the discussions during technical implementation of the GPT-based text summarizer and to editing the manuscript. A. Butz contributed two rounds of reviewing and editing of the manuscript.
Ou et al. [112]	D. Buschek contributed to the framing of the publication. S. Mayer contributed to the overall structure and clarification of the publication. A. Butz advised the research with close discussions over a year and contributed to four rounds of reviewing and editing of the manuscript.

Ou et al. [114]	S. Mayer contributed to the framing, structure, clarification of the publication, and one round of final editing of the manuscript. A. Butz advised the research and contributed to reviewing and three rounds of final editing of the manuscript.
Ou et al. [115]	Y. Zhan contributed to the user interface design of the user study, and Y. Chen contributed one round of the final editing of the manuscript.

Table 1: Clarification of contributions in specific projects and publications.

Table of Content

1
Introduction

It is beyond a doubt that all our knowledge begins with experience.

– Immanuel Kant, *Critique Of Pure Reason*, 1781

With the increasing interest in human-AI interaction, *human-in-the-loop (HITL)* [101] systems have been applied to a wide range of domains, such as material design [12], animation design [11], photo color enhancement [82], image restoration [155], and more [23, 53, 80, 163]. These systems actively exploit human choices to optimize machine results. They propose a set of design alternatives and then iteratively adapt their results based on choice preference feedback, thereby increasing the quality of the system outcomes and the satisfaction of the human involved while simultaneously speeding up the process.

One of the ancestors of these approaches is the Design Galleries approach [96]. 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. However, it is implicitly believed that the process would always converge and end in a desired solution. While following the assumptions, more and more reasons are discovered that challenge these assumptions, such as domain context [130], timing [51], trustworthiness [73], cognitive biases [12], unstable and contradicting preferences [112]. A human-in-the-loop system is not always beneficial to the user, and the system might not always be effective in achieving user satisfaction.

With these brief observations from the literature, this thesis was initially motivated by the successfully advertised human-in-the-loop system but towards presenting a set of empirical studies to investigate the boundary for human-in-the-loop optimization systems to be beneficial. In the interaction loop, machine agents are designed rationally to interact with human beings that may behave using incomplete rational policies iteratively.

1.1 Thesis Scope

Over the past few decades, computing system engineering has dreamed of a human-in-the-loop servo mechanism. A conscious human being is usually believed, in a rational manner, to operate, assist, and control the machine to achieve desired objectives. One of the very early uses of the terms "man-in-the-loop" [26] or "human-in-the-loop" appeared in the 1980s [159]: "*...we used a human in the loop. Commands that the system could not successfully parse but which seemed reasonable were reinterpreted by a hidden operator who recast them into valid commands without the subject's knowledge... By necessity, the operator made decisions quickly and frequently...*" The fundamental idea of these usage refers to a control loop between two entities where a human operates a machine system that may be easily broken.

Over time, with the increasing developments of hardware and software complexity, machine systems are more and more reliable, error-tolerant, and self-defensive. The overarching vision of "using a human in the loop" changes from continuous operation to occasional intervention. One of the interpretations is, the process of building layers of automation and operation complexity to existing machine systems accumulates collective rational system designers' and engineers' intelligence. For example, a machine system involves not only relevant software but also requires specialized hardware chips, reliable communication protocols, algorithms that are robust to noise and errors, and the ability to provide reproducible results when the same input is given so on. All these incremental developments in the machine world built up a vision of artificial intelligence, and the area was centered around perfect rationality.

It is precisely because of the rational assumption that apart from critical areas [49], in many cases, machine systems exceed human ability. Hence, the human-in-the-loop mechanism is being reshaped towards an extended vision that uses machine intelligence to assist individual human decision-making [145] or creativity [37] in the human-computer interaction (HCI) research field. For the support of decision-making, system design principles were heavily influenced by Simon [142] and his intelligence-design-choice model of decision-making. Instead, despite evidence and practical success of generative models[1], whether machine system generates creativity hence harbinger of art's demise is still an open question [38].

While there are rich HCI research publications that focus on designing user interfaces to combine with recent developments in computational intelligence, increasing the ability of machine systems has started to fulfill some of the expectations: delegate complex tasks to rational and reliable machine systems, inspect rich information to support a decision; generate content to support human creativity; and etc. Still, most user interface research focuses on specialized, targeted user groups, a pattern that uses user feedback to optimize machine systems – *human-in-the-loop optimization* – appears as a research gap: Can we design a machine system to adapt its internal objective according to individual user interactions? Will it align with the users' goals and achieve better outcomes?

Not surprisingly, achieving a successful system adaptation still relies on an underlying rationality assumption to individual users' interaction behavior. With the observations and developments in social science, the assumption is strongly challenged, and humans may not always be able to provide reliable responses for machines to learn and adapt. However, these boundary conditions are rarely investigated in the HCI research field. As one of the goals of HCI research is to design systems that consider the user perspective more if a system aims to adapt to individual users based on their inputs, how can we ensure users can provide reliable responses? How can we ensure the system can adapt to the user's expectations?

This thesis will focus on the *human-in-the-loop optimization* systems designed to optimize machine results based on human choice feedback. As previously discussed, we are particularly interested in the boundary for human-in-the-loop optimization systems to be beneficial. More broadly, how will the involved individual human intelligence fit into the collective machine intelligence in an interaction loop?

1.2 Research Questions

To fully understand the impact of the building blocks of human-in-the-loop optimization systems, this thesis identified six key areas. Figure 1.1 provides a visualized structure to the research questions and relevant chapters of this thesis.

[1] `https://openai.com/blog/chatgpt/`, *last accessed 17.02.2023*

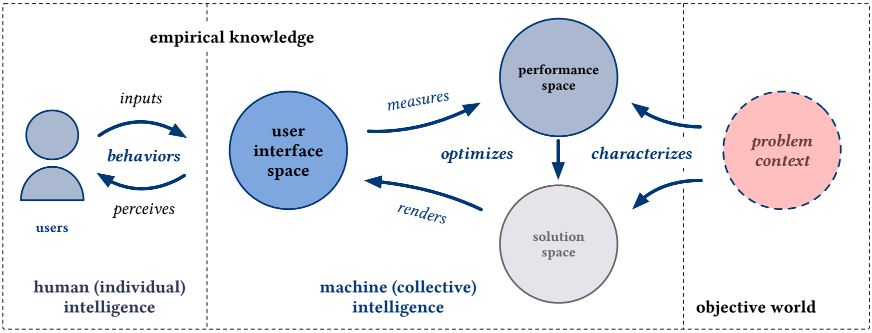


Figure 1.1: The intelligence in the human and machine interaction loop.

In this framework, under a specific problem context, the system ability or machine's built-in intelligence is built on top of collective intelligence from system designers and engineers. Such intelligence aims to solve a particular problem, constructs a space of solutions based on many design decisions when solving the problem, and all solutions can be explored using the exposed user interface. Based on the perception of the machine outcomes, the user uses their intelligence to provide feedback inputs back to the machine system through the user interface. Then, the machine will measure the overall system performance from different angles, and shape an observed performance of the current state, which is constructed as a performance space. The machine system will then optimize the performance space to find the best solution and adapt it to the user. The machine system will then present the new solution to the user, and the process repeats until termination criteria are met.

Therefore, with this overarching process, the guiding research questions of this thesis are:

RQ1 *Problem context.* What are suitable problem domains we should consider when using a human-in-the-loop strategy?

RQ2 *Performance metrics.* What are the relevant performance metrics, and how can we measure them to facilitate the performance comparison between human-in-the-loop systems?

RQ3 *User interface.* What are the current user interface design practices, and which interface suits human users better in the context of human-in-the-loop optimization?

RQ4 *Termination criteria.* What are the most suitable termination criteria, and how can the quality of human responses influence this criterion?

RQ5 *Human expertise.* How does the involved user expertise impact the system outcomes and subjective satisfaction?

RQ6 *Objective alignment.* How can we identify the alignment of objectives between the human user and the machine system?

We will discuss the answers to these research questions in the following chapters.

1.3 Contributing Publications and Outline

As a note, this monograph dissertation comprises multiple previously published results, including chapter 3-7. These results were published as individual research projects framed for better focus but fundamentally connected from project to project. This thesis presents an overarching picture and builds on top of the other relevant literature.

Furthermore, the thesis also includes additional insights as a follow-up complement to prior publications. Table 1 clarifies the original contributions of the author of this thesis and the contribution of others to the relevant publications.

[114] Changkun Ou, Sven Mayer, Andreas Butz. 2023. The Impact of Expertise in the Loop for Exploring Machine Rationality. In the 28th ACM Symposium on Intelligent User Interface (IUI '23). ACM, New York, NY, USA, 15 pages.

[112] Changkun Ou, Daniel Buschek, Sven Mayer, 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). ACM, New York, NY, USA, 11 pages. `doi:10.1145/3543758.3543761`.

[113] Changkun Ou, Sven Mayer, Daniel Buschek, and Andreas Butz. 2024. Rethinking Opinion Measurement Interfaces for Human-in-the-loop Optimization. Transactions on Computer-Human Interaction (ToCHI). ACM, New York, NY, USA, 28 pages. `SUBMITTED`.

[115] Changkun Ou, Yifei Zhan, Yaxi Chen. 2019. Identifying Malicious Players in GWAP-based Disaster Monitoring Crowdsourcing System. In the 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD). IEEE. New York, NY, USA, 10 pages. `doi:10.1109/ICAIBD.2019.8836972`.

[112] and [115] received Honourable Mention and Best Paper Awards at the respective conferences.

The research contributions can be classified following the definitions proposed by Wobbrock and Kientz [161], who introduced seven types of contributions, including artifact contributions, methodological contributions, and theoretical contributions. We follow with their description of the corresponding contribution.

Theoretical Contribution: This thesis presents an overarching framework for different building blocks of the human-in-the-loop system, which enables theoretical and empirical analysis of human-in-the-loop systems. In chapter 5, the thesis contributes a taxonomy space of user interfaces for measuring user opinions in human-in-the-loop systems; chapter 6 contributes a theoretical analysis of possible human and machine errors that can occur in a human-in-the-loop system; Lastly, chapter 7 contributes a theory and interpretations regarding the objective alignment between the user and the interacting machine.

Empirical and Artifact Contribution: The thesis discusses multiple empirical user studies designed to analyze human-in-the-loop systems' different building blocks. As open source artifacts[2], it also contributes three open-source real-world human-in-the-loop systems to actual human responses in text summarization, image color enhancement, and 3D model simplification problem contexts, including the dataset collected from the user studies conducted in this thesis. These datasets are open-sourced and can benefit much other future research on human priors and human-in-the-loop systems. Together, these contributions support researchers and developers in investigating and building future human-in-the-loop systems that are more robust and more beneficial.

Methodological Contribution: The thesis presents a comprehensive recording and analyzing methodology for human-in-the-loop systems. It deliberates the necessary aspects to measure when analyzing a human-in-the-loop system and a common experiment framework to common abstract procedures of user studies regarding human-in-the-loop systems.

This thesis is structured as follows:

Chapter 1: Introduction introduces the motivation of this thesis and an overview of the research questions and this outline.

[2] `https://changkun.de/s/intelligence-in-the-loop`, *last accessed 17.02.2023*

Chapter 2: Background and Definitions comprises the related work and the usage of terms in human-in-the-loop systems, approaches for inferring preference from human feedback and better adaptation, psychological theories of the rationality of humans and judgments, and the understanding of human expertise and approaches to measure it.

Chapter 3: Problem, Solution, and Performance Spaces presents the problem, solution, and performance spaces, which are the fundamental building blocks of human-in-the-loop systems on the machine side. The chapter discusses the suitable problem domains and selection approach used in this thesis, then presents the existing objective metrics to measure the solution space. This chapter addresses RQ1 and RQ2.

Chapter 4: Experiment Design and Apparatus deliberates a common experiment abstraction across all other chapters that appeared in this thesis and also discuss the necessary engineering details for implementing the systems being used in the user studies.

Chapter 5: Measuring Opinions with Interfaces investigates a taxonomy of opinion measurement user interfaces for human-in-the-loop systems and conducted a study to inspect the impact of different user interfaces on the quality of human responses. This chapter addresses RQ3.

Chapter 6: Termination Condition look closer into the termination criteria of human-in-the-loop systems, especially from the human perspective. The chapter describes two possible error sources that may be overlooked while designing human-in-the-loop systems and demonstrates the possible solutions to mitigate them. This chapter addresses RQ4.

Chapter 7: Expertise and Objective Alignment addresses how involved human expertise can impact system outcomes and subjective satisfaction. Based on the results, the chapter also discusses possible interpretations of aligning human objectives and the internal machine optimization process. This chapter addresses RQ5 and RQ6.

Chapter 8: Reflections and Outlook reflects on the work presented in this thesis. Furthermore, it includes several reflections on our observations, the directions for future research, and how they can benefit from this work.

2
Background and Definitions

Every age has its myths and calls them higher truths.

– Morris Kline, *Mathematics: the loss of certainty,* 1982

Before diving into the main contributions of this thesis, this chapter[1] provides a brief overview of the related work and the main concepts that are used throughout the thesis in 1) recent advances in *human-in-the-loop optimization* systems; 2) machine learning approaches, especially *Bayesian optimization* regarding inferring human preference from choice; 3) cognitive and social psychology concepts and theories regarding *rationality* and *satisficing*.

2.1 Human-in-the-loop Systems

The human-in-the-loop strategy may be applied in different domain contexts, which connect to different design goals, including personalization, co-creation, and decision-making support. Primarily, concerning the human-in-the-loop strategy, it is either using humans as a servo mechanism to steer pre-defined machine behavior or utilizing machine algorithms to analyze and optimize the machine behavior based on human inputs.

[1] The content of the chapter is partly based on Ou et al. [112, 113, 114].

From the human computation [123, 151] perspective, a human-in-the-loop system may be designed to use crowds [56] 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 [66], there are several identified challenges [127] 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 [85], 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, human-in-the-loop systems using crowds may suffer from malicious inputs [115] 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 [80] is considered limited when individual customization has a higher priority.

In a personalized context, Buschek et al. [14] 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* [89], and *usability* [77]. They highlighted a trained AI contains system bias, but lacks discussions on where is the source of bias and how much mismatch between individual expectations and system abilities. In terms of mismatched expectations, Eiband et al. [36] 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. [150] 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 user 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 [16, 74]. Still, it is *implicitly assumed* that the human involved eventually makes a rational decision over *subjectively* untrusted AI outcomes. Factors such as algorithm aversion [29] were confirmed to indicate that users are more biased [124] towards human results and produce considerable noise even in the judicial area [31].

Although prior research [23, 53, 80, 96, 155, 162] that involves human-in-the-loop strategies have shown human knowledge to be helpful for a machine to learn, previous literature rarely discusses the circumstances under which human-in-the-loop 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.

This thesis is an attempt to address this research gap by analyzing different building blocks of a human-in-the-loop system and then showing practical challenges that can arise to align a human-in-the-loop system to the user when exploiting their individual preferences.

2.2 Inferring 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, this thesis follow Hausman [50] in their counterargument against eliminating preference using choice [48] and acknowledge the existence of *preference.* In our use of terms, shown in Figure 2.1, 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 [146] 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 one can infer that A is preferred over C if A is preferred over B and B is preferred over C.

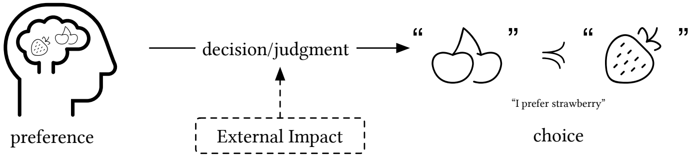


Figure 2.1: A decision process turns an internal preference state into an observable choice. The choice may not reflect the underlying preference due to external influences.

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 [130], and the transitivity axiom is not applicable when implicitly involving other judgments that were not previously considered. For example, when a human decides to use two objectives and chooses A over B and B over C involving only one objective, they may implicitly involve the 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*. For obtaining preference from choice, a learning approach has to effectively deal with this uncertainty regarding the presence of preference and the rationality of an individual being.

Bayesian Optimization (BO) [102] provides a way to model human feedback [21, 40] and enables human-in-the-loop optimization [137] by utilizing, e.g., biological measurements [30, 72], interaction behavior [19], and measured preferential choices [12, 45, 83]. It learns a *posterior* from the measured human feedback and aims to search for a global maximum of an unknown function by exploration and exploitation. In conjunction with measured human feedback, two popular genres of BO are actively being researched, as shown in Table 2.1.

Since Utility-based BO [102, 133] requires measuring human feedback as a utility value and learns a *posterior* from choice, it aims to search a global maximum of an unknown function by exploration and exploitation. Therefore, it can be used to propose examples using an *acquisition function*, ask the human to provide a choice, and then infer the underlying preference iteratively. In contrast, When dealing with choices from pairwise comparisons, preference-based BO (PBO)[2] can propose examples using an acquisition function, ask the human to provide a choice and infer the underlying preference iteratively. As a specialized category of BO, PBO has received increasing development in recent studies [45, 82, 100, 140]. 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 [146].

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 [148] have widely presented how heuristic biases might influence the choice behavior. Ex-

[2] This thesis 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. [45].

Table 2.1: Examples of different Bayesian Optimization frameworks. Types of feedback can be either utility-based or preference-based. Most of the BO frameworks do not model indifference, and none of them consider incomplete preferences. Note that 1) rating can be considered either as absolute utility or preferential ranking distance depending on the underlying preference model, 2) selection is a special case of ranking where one item is strictly preferred over other options, and the ranking order of those unselected options is not specified, which means that their preference relation may be considered incomplete.

Optimizer	**Input**	**Feedback**	**Options**	**Indiff./Incomp.**
Naive BO [102]	Rating	Utility	Pointwise	No/No
2GC [12]	Selection	Preferential	Pairwise	No/No
PBO [45]	Selection	Preferential	Pairwise	No/No
EUBO [93]	Selection	Preferential	Pairwise	No/No
4GC [11]	Rating	Preferential	Listwise	No/No
SLS [83]	Selection	Preferential	Listwise	No/No
SPS [82]	Selection	Preferential	Listwise	No/No
PPBO [100]	Selection	Preferential	Listwise	No/No
Top-k BO [105]	Ranking	Preferential	Listwise	Yes/No

ternal causes can also produce a considerable amount of noise in choice [62].

To avoid the mentioned violations of the transitivity axiom, the recent extensions [6, 82, 140] 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 [6, 82]; or provide a ranking of all given options, where options may share the same level of rank [140].

To overcome the violation of preference axioms, PBO has also considered handling noisy inputs [91] and guarantees theoretical convergence when dealing with unstable choices. Despite all these developments in PBO, there still exist many 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 is capable of dealing 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 to design human-in-the-loop systems carefully, one must design the UI carefully to mitigate these issues and reduce user errors. Note that more ranking elements may also increase the uncertainty for users to make imperfect decisions [101] due to increased workload. Thus, one should carefully consider the presented number of elements.

2.3 Human Rationality and Judgments

Apart from inferring human feedback using a learning algorithm, 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 [141] first coined the term *bounded rationality* to describe the perceived information limits of individual rationality. This observation provides a sufficient discussion base for interpretations regarding irrational decisions. As previously discussed, Tversky and Kahneman [148] emphasized one possible category of systematic errors from the cognitive perspective. In recent discussions [10, 44], researchers take a statistical perspective and underline that recurring noise could also contribute equally [62] 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 [132, 141] 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. [135] 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 another example, prior work [57] showed that a satisficing strategy leads to the quicker selection, hence increasing 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 human-in-the-loop 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, it is also interesting to understand 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 [47], which proved less efficient and more correlated to a maximizing strategy. Still, it remains unclear what the satisfaction would be in this case.

Economics widely studied decision-making when choosing a preferred item from several alternatives. The expected choice utility maximization [97] 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 [141] 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* [64, 149] 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 [148] showed that in any decision under uncertainty, System 1 (fast and instinctive thinking) tends to override System 2 (slow and rational reasoning) [60], 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* [138] may occur as wealth increases and marginal utility decreases; 2) *loss aversion* [63], as part of the *endowment effect* [61], 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 [10, 44], systematic noise descriptively shapes another form of decision error that contributes equally to judgment error as individual bias [62]. The decision noise components [62] break down into *level noise* (decision variability between groups), *stable pattern noise* (contextual bias within groups), and *transient noise* (purely occasional).

2.4 Expertise and Intelligence

Different than context-dependent rationality and satisficing, the involved expertise of a human, i.e., the intrinsic intelligence and knowledge of a human, is constructed in their long-term experiences and is considered more effective in the interaction and decision process. Therefore, to further understand the human factors in the context of human-in-the-loop optimization, it is also important to understand the concept of expertise and its impact on the interaction and decision process.

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. [43] describe six different dimensions regarding expertise, whereas Collins [24] suggests three dimensions and Kotzee and Smit [79] suggests only two dimensions based on social aspects. On a higher level, Bourne Jr. 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 [69, 70].

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 [86]. For our purpose here, it is interested in quantifying the level of expertise of a specific human within a particular context. Treem and Leonardi [147] 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 [111] 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 [90]. Ferrod et al. [39] 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. To measure the involved expertise in a feedback loop, one does not only need to measure the accumulated human experience but also needs to consider the context involved.

2.5 Summary

This chapter shaped a few fundamental terminologies regarding the initial understanding of the human-in-the-loop system. Although the definition of the term "human-in-the-loop" is loosely different from "interaction" itself, with a parse of recent literature, human-in-the-loop related research mainly focuses on the direction of humans takes the lead, and machine analyzing and learning user inputs progressively so that the background machine system can produce unique adapted content, which led to an optimization context. To achieve this goal, there are many existing practices in the field of machine learning. While in the scope of this thesis, we mainly looked into the literature regarding using *Bayesian optimization* to model users' preferences based on their decision-choice behavior; hence, we will also use "human-in-the-loop optimization“ to refer to such a machine system. With a connection to an actual human, we also inspected literature that formerly investigated the imperfection of human decisions and the concept of expertise. All these backgrounds will be further discussed in the following chapters when presenting our empirical explorations and reflections.

3
Problem, Solution, and Performance Spaces

Existence precedes and rules essence.

– Jean-Paul Sartre

To start, this chapter[1] will elaborate on the problem space, solution space, and performance space to facilitate the empirical explorations in this thesis. We will discuss the problem space in detail, including our selection approach of the problem contexts and the corresponding solution spaces of the related algorithms. We will discuss possibilities to measure the performance regarding the overall behaviors of the system between user, machine, and underlying optimization approach that aligns machine behavior towards an individual human.

3.1 Problem and Solution Space

In a human-in-the-loop optimization scope, 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 strictly convex continuous function). We need to select tasks where users provide ranking feedback using their

[1] The content of the chapter is partly based on Ou et al. [113, 114].

expertise. A task should also be iterative for observing progress and partially subjective because users could balance the trade-off on different objectives.

With the above consideration, this thesis 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 fundamental human expertise; photo color enhancement involves 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 human-in-the-loop optimization literature [81, 82, 83, 112, 143] individually but not compared to each other together.

In order to minimize the problem scale, in chapter 5, we will discuss the user interface design for the text summarization and photo color enhancement design space. In chapter 6, we will look closer at the 3D model simplification design space. Lastly, in chapter 7, we will conduct a user study and compare across all three contexts.

Below, we will first discuss the problem and solution space in detail for each domain.

Domain 1: Text Summarization In the domain of text generation, text summarization is the problem of generating a short summary of a given text. It is a well-studied problem in natural language processing (NLP) and has been used in many applications, such as news summarization, document summarization, and query summarization. When dealing with text summarization, the general approach can be considered in two directions [87]: 1) Extract the most important sentences from the original text, and pick the parts of the text with the highest relevance defined by a correlation distance. The selection process directly produces the summary. 2) Generate a summary from the extracted sentences, similar to what most humans would do when tasked to summarize a text.

The first direction is called *extractive summarization*, and the second direction is called *abstractive summarization*. The former is a simpler task and can be solved by a greedy algorithm [99], while the latter is a more challenging task and requires a more sophisticated model. As a mix of both approaches, See et al. [136] allows a model to replicate some parts verbatim while rewording the others.

Domain 2: Image Color Enhancement In contrast to many other disciplines, the issue of picture color enhancement is not clearly defined, and applicable solutions are often evaluated based on basic heuristics. One of the most important reasons is individual differences in color perception and prior education [17]. This allows us to automate the image quality enhancement process by using machine learning techniques [15] to understand the connection between many images and, depending on a heuristic measure and preferential feedback [81], the underlying goal of machine algorithm is to maximize the appearance of an image based on the feedback from a user.

Therefore, when the user teaches the algorithm by manually providing feedback, the algorithm tunes parameters that alert the color filters to alert the overall artistic style of an image. In the majority of cases, the user is not aware of the underlying algorithm, but the algorithm exposes a set of explainable features to the user, which can be used to understand the overall direction of a user to improve the image color without being aware of all underlying details.

Domain 3: 3D Mesh Simplification For the specific geometry processing problem, polygon reduction, presented in this thesis, we briefly discuss its recent advances and the relevant methods we utilized to build our system. The geometry processing "No-Free-Lunch" theorem [154] 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 [41, 42] and have also been used for levels of detail (LOD) generation [54], even baked into hardware rendering pipelines [110]. They are efficient but contain ill-posed cases with results depending on the implementation. Instead, the general idea of global remeshing [7, 35, 76, 120] 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. [68], 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 [55, 58] and local [42, 54] methods, but this also introduces the new problem of cutting a mesh. We refer to Metis [71, 121] as a mature solution for graph partitioning.

3.2 Performance Space

As shown in Figure 1.1, the problem context and the defined solution space together will produce system outcomes for the user, then combined with the user feedback, a set of considered measures defines the overall system performance.

To facilitate the inspection and analysis of the performance of human-in-the-loop systems, one can consider an interaction loop more successful from the human side if 1) the involved human spends less time on providing feedback inputs, 2) the user can achieve a satisfactory result in fewer interactions, and 3) the user interface allows the human to give more precise and clear feedback. From the machine side, the underlying optimizer can learn human feedback more effectively and propose outcomes of high objective quality faster. In this subsection, we will discuss the considered metrics for the human side and the machine side in this thesis and how they are measured in subsequent studies.

3.2.1 User Performance

The performance of a user in a human-in-the-loop system can be considered from two different aspects: behavior and responses. The former is related to the overall interaction with the system, while the latter is related to the user's feedback to the system. They are two different channels of information that can be used to evaluate and verify each other.

Behavior Measures Behavior can be considered from different perspectives. From the user's interaction perspective, one can measure the overall interaction of the developed system, including 1) decision time, starting from when the evaluation options were presented to the last time when users interacted with the interface; 2) occurrence of incomplete preferences per participant per iteration, 3) occurrence of expressing indifference 4) the number of mouse operations to select/rank the given system outcomes in listwise interfaces.

User Responses To evaluate human responses while interacting with a human-in-the-loop system, one can collect rating and preferential ranking data in each iteration of interaction as user responses. One can use the utility value from direct rating or the learned latent utility from preferential rankings to quantify this user feedback. Besides the responses collected while interacting with the system, there are two additional measures regarding expertise and overall satisfaction.

Before participation, a user's demographics are measured, especially background knowledge and expertise. As discussed in section 2.4, since user expertise is measured differently in prior research, we use a similar approach as Ferrod et al. [39], Ooge and Verbert [111], and combine the following established metrics: 1) self-indication, 2) accumulated work experience, and 3) recent experiences in the domain. For the text context, we ask for their language proficiency; 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 most recent 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 users, then grouped users into three groups using quantile-based discretization: *novice*, *intermediate*, and *experienced*. Note that the descriptions only represent the relative levels among our recruited users. In a larger user group, they may be reconsidered as novice or intermediate accordingly.

Lastly, at the end of participation, there are six questions presented to measure the overall subjective satisfaction regarding the machine outcome: Q1) participants' overall subjective satisfaction with the final results; Q2) the confidence they think they can do a better summarization by themselves than the system, which was optimized based on their provided feedback; Q3) whether the outcome matched their expected summary; Q4) whether they felt improvements of the summarization from iteration to iteration; Q5) whether they felt the "I don't know" option was helpful, 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.

3.2.2 Domain-specific Objective Outcome Measures

For the performance on the machine side, there are interior and exterior measures. The interior measures are related to the Bayesian optimizer's behavior, while the exterior measures are related to the system outcome and corresponding domain-specific metrics regarding the objective quality.

Bayesian Optimizer To measure the behavior of a Bayesian optimizer, one can consider using prior input information that maps a set of collected user responses to a set of optimized parameters. Based on the estimated posterior, one can measure the expected improvement of the optimizer's performance in the next iteration. This enables us to understand the optimizer's learning progress and the system's overall performance.

Text Summarization There are many different approaches to define the overall similarity between two paragraphs. The core idea of the existing measurement is to compare a machine-generated text to one or multiple human-generated texts. This type of comparison might be biased by the human's writing style. Therefore, the comparison also often aggregates a large number of human results. In particular to the summarization context, the total number of words in the outcome texts to validate if a user made progress on the given objectives. We can also compute the BLEU [117] and ROUGE [92] (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.

Image Color Enhancement For the image color enhancement, we can compute 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 can compute the mean pixel-wise difference of U channels in YUV color space for tint changes and similarly in the V channel for temperature changes. While these measurements do not reflect actual image "quality", together with the strategy of the optimizer to always show the current best choice, a consistent trend here can be seen as an indication of continued improvements. Although those metrics stay at a technical level, they positively correlate with human judgment.

3D Model Simplification 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 [25]. Therefore, we can compute the simplification ratio to validate whether users 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 can use an equally weighted combination of *peak signal-to-noise ratio* [78] (PSNR), and *structural similarity* [153] (SSIM) to measure the rendering difference. For wireframe quality, we can compute *scaled Jacobian cell quality* [75] because it was previously suggested to better correlate with the human judgment [25]. The scaled Jacobian cell quantity itself measures how a given face is regularized. Lastly, we can sample two point clouds on the source mesh and the outcome, then use *Chamfer distance* [4] to indicate the surface quality. Surface quality is less observable compared to the other three objectives.

3.3 Summary

This chapter discussed the relevant measurements in a human-in-the-loop optimization system, including human-side expertise and satisfaction measurement, behavior measurement such as human choice feedback, interaction behavior, timing, and involved iterations. We also discussed the domain-specific objective outcome measures for text summarization, image color enhancement, and 3D model simplification, which will be used in the following chapters for analyzing the overall system outcome quality. The measurement using these metrics contributed to the shape of the overall performance space; they will also facilitate the verification of different observation channels and are beneficial for the overall discussion of the interpretation of the results.

4
Experiment Design and Apparatus

There are two methods in software design. One is to make the program so simple there are obviously no errors. The other is to make it so complicated there are no obvious errors.

– Tony Hoare

In previous chapters, we discussed the relevant background knowledge and the considered problem contexts in this thesis. Before we narrow down to the specific problem domains, this chaper[1] demonstrates a general abstraction of the experimental design throughout the thesis and the engineering details of relevant domain-specific systems.

4.1 Design Patterns of Experiment

For a human-in-the-loop optimization system, the experiment that evaluates how users interact with it will involve different parts: 1) understanding the user background, 2) sampling and calibrating individual differences, 3) interaction and optimization loop until termination criteria are met, and 4) collecting user feedback and experience. Therefore, the overall four experiment stages for a participant are visualized in Figure 4.1, and the different phases are explained below.

[1] The content of the chapter is partly based on Ou et al. [112, 113, 114].

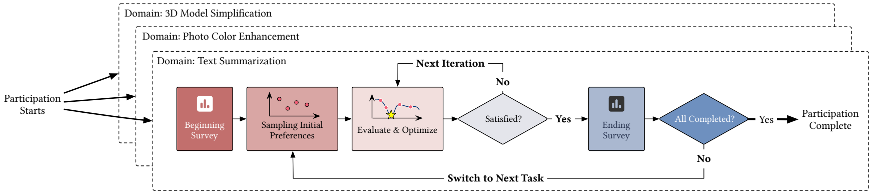


Figure 4.1: A general procedure of all studies discussed in this thesis. A participating user conducts one of the problem domains: text summarization, photo color enhancement, and 3D model simplification. The experiments reported in chapter 5 explores the domain of text summarization and photo color enhancement. chapter 6 mainly focuses on the 3D model simplification. An overall comparison across three domains is reported in chapter 7.

Study Intro Participants started participation with an informed consent form and answered initial demographic questions, including their age, gender, and levels of expertise.

Main Task The user interface presents a set of machine outcomes to participants. The main task for a user is to provide feedback to the machine using the UI. Based on the collected user feedback, the background optimizer will iteratively optimize the machine outcomes. Because the optimizers need initialization samples to fit individuals, to minimize the participation time, in the first four iterations, a participant evaluated outcomes produced by quasi-random Sobol sampled [128, 158] system parameters. After acquiring these preference priors from participants, starting from the 5th iteration, the optimizer is used, and participants can freely terminate when satisfied with the current system outcomes. The task ends automatically after 20 iterations to limit participation time.

In particular to the problem domains discussed in this thesis, we use the following three main tasks as empirical examples:

- *Text summarization*: Given a news article, the system produces a summary of the article. Based on the user feedback, the system optimizes the summary to be more concise while preserving the meaning of the article.

- *Photo color enhancement*: Given a photo, the system produces a color-enhanced version of the photo. Based on the user feedback, the system optimizes the color-enhanced photo to be more pleasing to the user.

- *3D model simplification*: Given a 3D model, the system produces a simplified version of the model. Based on the user feedback, the system optimizes the simplified model to be more pleasing to the user.

Ending a Task After termination, participants answered six questions regarding their satisfaction with the system outcomes and their experience with using the system to give feedback.

Ending the Study Each participant will need to complete Latin square-shuffled machine outcomes. Specifically, articles in text summarization tasks[2], photos in image color enhancement tasks[3], and 3D models in mesh simplification tasks[4].

4.2 Apparatus and Engineering

To successfully execute the experiment, we must build a system for participants. In this section, we will discuss high-level engineering details of the system and the infrastructure that we used to run the experiment.

For the frontend interfaces, there are potentially two different kinds of needs. In the lab experiment context, the user interfaces must be controlled, and the overall experiment procedure must be managed to minimize confounding variables. In the wild experiment context, the user interfaces must be easy to use and provide a good user experience. In both cases, a possible reuse of the facility is to separate the core functionalities into the backend so that the front end can be easily replaced by different UIs as long as the backend APIs are compatible.

In the subsequently presented user studies, we developed the frontend web UIs using Material UI[5], React DnD[6], and three.js[7]. For the native desktop software UI, we collaborated with an industrial partner[8], who useed Unity[9].

[2] Selected from the CNN daily mail dataset, article IDs: ea06fd0b, 35f0e33d, 42c027e4. See `https://huggingface.co/datasets/cnn_dailymail`, *last accessed 17.02.2023*

[3] Selected from Koyama et al. [82]. See `https://github.com/yuki-koyama/sequential-gallery/tree/main/resources/scaled`, *last accessed 17.02.2023*

[4] Selected model name: Stanford bunny, Suzanne, and fan disk. See `https://github.com/alecjacobson/common-3d-test-models`, *last accessed 17.02.2023*

[5] `https://mui.com/`, *last accessed 17.02.2023*

[6] `https://react-dnd.github.io/react-dnd/about`, *last accessed 17.02.2023*

[7] `https://threejs.org`, *last accessed 17.02.2023*

[8] Ingman Technologies GmbH and WAY Engineering GmbH.

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

Apart from the frontend, We expose the core functionalities of our system as Web APIs that run the different computation services, such as text summarizer, image color enhancer, and mesh reducer. The backend *core service* is written in Go[10] for easier concurrency management. It serves all frontend interfaces, data collection, and communications with 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. All services are deployed on the institute infrastructure (Ubuntu 20.04, 8-Core Intel Core i9-9900K, 64GB RAM, and NVIDIA GeForce RTX 2080 Ti with 11GB of GPU memory).

We implemented three separate domain services. To perform text summarization, we picked the pre-trained BART model via HuggingFace[11]. We implemented an isolated text summarization server using Flask[12] with GPU acceleration. We use Nucleus sampling [52] as a stochastic text decoding strategy[13] 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 parameters at every model inference stage: 1) *summarization ratio*, 2) *length penalty*, 3) *top-p*[13], and 4) *temperature*[13].

For photo color enhancement, we used a parameterized photo enhancer [81, 82, 83] as an image processing service for better integration to the core service. This service allows five adjustable system parameters that are bounded between 0 and 1: 1) *brightness*, 2) *contrast*, 3) *saturation*, 4) *temperature*, and 5) *tint*. Lastly, we engineered a hybrid local/global algorithm based on state-of-the-art research [42, 58], as a 3D mesh processing service, and it also contains five bounded system parameters: 1) *simplification ratio*, 2) *border preservation*, 3) *hard edge preservation*, 4) *sharpness preservation*, and 5) *quadrilateral preservation*.

We implemented the underlying optimizer using BoTorch [3] as a command line service, which reads the user responses to estimate the next optimal system parameters for exploration. BoTorch provides the EUBO [93] optimizer as one of the state-of-the-art PBO optimizers that consider noisy inputs to estimate system 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

[10] `https://go.dev`, *last accessed 17.02.2023*

[11] `https://huggingface.co/sshleifer/distilbart-cnn-12-6`, *last accessed 17.02.2023*

[12] `https://flask.palletsprojects.com/en/2.2.x/`, *last accessed 17.02.2023*

[13] `https://huggingface.co/blog/how-to-generate`, *last accessed 17.02.2023*

the learned latent utility value to fit another Gaussian process and infer the next batch of exploration positions. The acquisition function is the Analytic Expected Utility Of Best Option acquisition function [93]. Typical extensions of BO model a preference function that transfers the problem of learning a utility function to learn a preference function that implies a latent utility function, e.g., González et al. [45]. However, EUBO models the expected utility of system outcomes based on a given user response dataset $\mathcal{P}_n = \{(\mathbf{x}_{1,i}, \mathbf{x}_{2,i}, r(\mathbf{x}_{1,i}, \mathbf{x}_{2,i}))\}_{i=1}^{n}$ where $\mathbf{x}_{1,i}$ and $\mathbf{x}_{2,i}$ are the pair of interests, and $r(\cdot)$ is the response of a user that indicates either 1 prefer the first outcome of the pair or 2 refers to the second outcome. The acquisition function of EUBO considers maximizing the *expected outcome utility difference* between iterations $\mathbf{E}_i\left[\text{argmax}_{\mathbf{x}\in\mathcal{X}}\, \mathbf{E}_{i+1}[f(\mathbf{x})] - \text{argmax}_{\mathbf{x}\in\mathcal{X}}\, \mathbf{E}_i[f(\mathbf{x})]\right]$ where f is the assumed latent utility function[14]. The design of the EUBO guarantees that the current optimum is included in pairwise or listwise comparisons, and we can use maximum posterior expected utility [21] to understand the impact on outcome utility.

For utility-based optimization context, we implemented a straightforward utility-based BO; for preference-based UIs, we use EUBO [93] that is designed for pairwise comparisons; and for all listwise ranking context, we adopted EUBO such that the optimizer can infer a batch of outcomes based on ranking preferences.

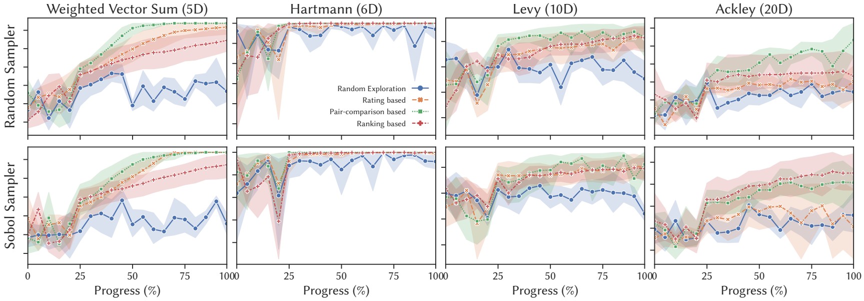


Figure 4.2: The performance benchmark for validating implemented Bayesian optimizers compared to a random exploration (baseline). Higher y-axis utility values mean better exploitation performance. The upper row shows exploitation performance using a random sampler for the initial samples. The bottom row shows exploitation performance using a Sobol sampler. Each column shows the performance of a different synthetic function. The results validate our implemented BOs and outperform random exploration.

[14] This acquisition function is intractable due to nested structure, but Lin et al. [93] showed a simpler equivalent $\mathbf{E}_n\left[\text{argmax}\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}\right]$ that preserves the same design property.

As a demonstration of the validity of the implemented Bayesian optimizer, we selected four standard synthetic functions[15]: Weighted Vector Sum (5D), Hartmann (6D), Levy (10D), and Ackley (20D) for benchmarks. The results demonstrate that all of our implemented Bayesian optimizers outperform random exploration.

4.3 Summary

This chapter introduced an abstraction of human-in-the-loop optimization experiment design patterns. We presented how three considered problem contexts fit into the same experimental procedure. We also discussed a detailed implementation of the experiment procedure regarding text summarization, photo color enhancement, and 3D model simplification. In the next chapters, we present experiments based on these details and discuss the findings from those empirical explorations. Specifically, in chapter 5, we will discuss the user interface design for the text summarization and photo color enhancement design space. In chapter 6, we will look closer at the 3D model simplification design space. Lastly, in chapter 7, we will show a user study and compare all three contexts.

[15] See `https://www.sfu.ca/~ssurjano/optimization.html`, *last accessed 17.02.2023*

5

Measuring Opinions with Interfaces

Divide each difficulty into as many parts as is feasible and necessary to resolve it.

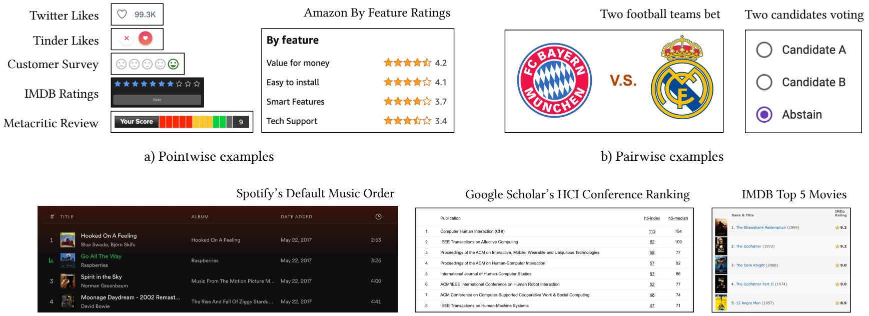
– René Descartes

This chapter[1] revisits the design space of opinion measurement interfaces in the context of human-in-the-loop optimization and elaborates on the terminologies used to differentiate user interface variants. Moreover, the content discusses an experiment to validate four major structured hypotheses. It shows the impact of user interfaces on a human-in-the-loop optimization loop by comparing multiple design variations of opinion measurement interfaces. The overall results suggest two optimal interfaces: 1) pairwise non-forced choice interface; or 2) listwise utility and preferential choice hybrid interface. The choice between the two is a tradeoff between decision time, the precision of measured human feedback, and overall optimization loop performance.

5.1 Opinion Measurement Interfaces

Conventional opinion measurement interfaces have different absolute measures, such as categorical choice and ordinal Likert-scale measures. The underlying

[1] The content of the chapter is partly based on Ou et al. [113].



By feature

Value for money	4.2
Easy to install	4.1
Smart Features	3.7
Tech Support	3.4

○ Candidate A
○ Candidate B
● Abstain

a) Pointwise examples

b) Pairwise examples

#	TITLE	ALBUM	DATE ADDED	
1	Hooked On A Feeling Blue Swede, Björn Skifs	Hooked On A Feeling	May 22, 2017	2:53
	Go All The Way Raspberries	Raspberries	May 22, 2017	3:25
3	Spirit in the Sky Norman Greenbaum	Music From The Motion Picture M...	May 22, 2017	4:00
4	Moonage Daydream - 2002 Remast... David Bowie	The Rise And Fall Of Ziggy Stardu...	May 22, 2017	4:41

	Publication	h5-index	h5-median
1.	Computer Human Interaction (CHI)	113	154
2.	IEEE Transactions on Affective Computing	62	109
3.	Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies	58	77
4.	Proceedings of the ACM on Human-Computer Interaction	57	92
5.	International Journal of Human-Computer Studies	57	86
6.	ACM/IEEE International Conference on Human Robot Interaction	52	77
7.	ACM Conference on Computer-Supported Cooperative Work & Social Computing	48	74
8.	IEEE Transactions on Human-Machine Systems	47	71

Rank & Title	IMDb Rating
1. The Shawshank Redemption (1994)	9.2
2. The Godfather (1972)	9.2
3. The Dark Knight (2008)	9.0
4. The Godfather Part II (1974)	9.0
5. 12 Angry Men (1957)	8.9

c) Listwise examples

Figure 5.1: Real-world measured opinions in different interfaces: a) Commonly used rating scales to measure opinion regarding one item; b) Two alternative candidates (non-)forced choice; c) Music album song orders without specified criteria, and forced ranking of conferences and movies using a distance.

analysis is based on Thurstone's law of comparison [146]. The optimal design of a pointwise measuring UI has been explored in psychology to reduce measuring error, for example, by asking repeated questions and checking the overall consistency [139], regarding the influence of ticks [98], or by allowing users to express uncertainty about their own input [46]. Previous work also discussed improving rating accuracy in the human-in-the-loop context [33, 106].

A first systematic evaluation regarding the design space was provided by Nobarany et al. [109]. They also observed that the rating interface might not be ideal for measuring opinions and discussed how to support ranking in the interface. In a way, this is similar to how PBO was designed [11, 12, 45]; opinion measurement interfaces that support comparison were also found beneficial here. Kalloori et al. [65] discussed a few reasons for better-measured opinions using preference-based feedback: 1) Explicit rating utility suffers from calibration of drift, for concepts with different types of interpretations might lead to differently rated results. 2) Users might change their opinion from time to time due to anchoring [119]. 3) Relative questions might potentially be faster and easier to answer. Due to the increased complexity of feedback types, ranking UIs are rarely explored, although there is initial work in this direction [95, 104].

Interfaces used in the Real World Opinions are frequently measured in real life for a variety of purposes. One can generally observe three genres of these UIs: pointwise, pairwise, and listwise UIs, as shown in Figure 5.1.

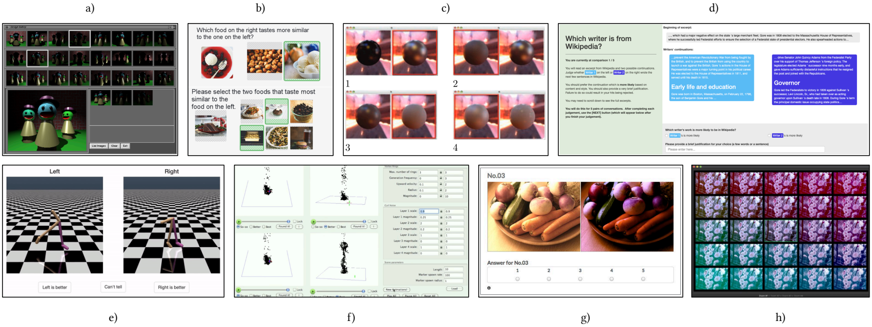

a)
b)
c)
d)
Which food on the right tastes more similar to the one on the left?
Please select the two foods that taste most similar to the food on the left.
1
2
3
4
Which writer is from Wikipedia?
Early life and education
Governor
Left
Right
Left is better
Can't tell
Right is better
No.03
Answer for No.03
1
2
3
4
5
e)
f)
g)
h)

Figure 5.2: Examples of opinion measurement interfaces being used in human-in-the-loop optimizations: **a)** A user chooses from a list of candidates to optimize illumination [96]; **b)** A user is asked to choose items that are closer to a given reference [157]; **c)** A user is required to decide which of the two given options is closer to their desired material [12]; **d)** A user is asked to choose which of two presented texts is more likely to be written on Wikipedia by a human [156]; **e)** A user is required to choose which trained machine's back-flip is better [20], note that the interface uses a "can't tell"; **f)** A user rates four presented graphics animations using a 3-step rating scale to show how close an animation is to what they are looking for [11]; **g)** A user rates which of the given paired designs is more preferred [80]; **h)** A user is required to select one target out of a list of choice alternatives [82].

Interfaces used in Human-in-the-loop Optimization As shown in Figure 5.2, there are many designs for opinion measurement interfaces used in human-in-the-loop optimization. Specifically, the presented examples can be treated as a mix of pairwise and listwise comparisons, and the underlying feedback type that models the UI's input form is either utility- or preference-based. Because of this diversity, providing a taxonomy is helpful in categorizing them and determining the study design in the next section.

5.2 Design Space and A Taxonomy

Table 5.1 provides an overview of the design space of opinion measurement interfaces in terms of the number of judging options and the types of feedback. The design space formally contains these two dimensions: the number of *alternative options* and the underlying *feedback type*.

Table 5.1: A design space of opinion measurement interfaces concerning feedback type, ordering type, and batch size. The numbers in bold are those of the interfaces implemented in the study. All of them allow users to express incomplete preferences using an "I don't know" checkbox.

	Alternative Options ($n \in \mathbb{N}$)		
Feedback Type	**Pointwise ($n = 1$)**	**Pairwise ($n = 2$)**	**Listwise ($n > 2$)**
Utility	1-RS (**UI1**)	2-RS	n-RS (**UI3**)
Preferential (strict)	N/A	2-AFC	n-AFR (**UI4**)
Preferential (weak)	N/A	2-ANFC (**UI2**)	n-ANFR (**UI5**)
Hybrid (strict)	N/A	N/A	n-AFRD
Hybrid (weak)	N/A	N/A	n-ANFRD (**UI6**)

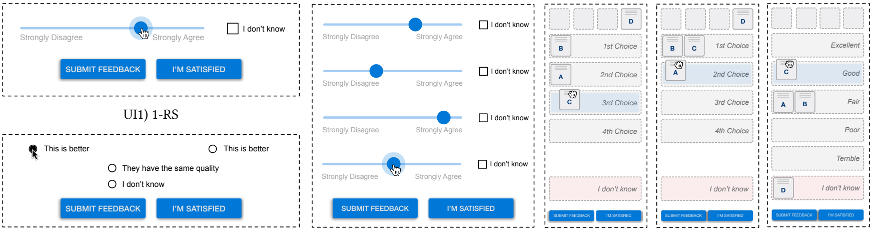
The number of alternative options encodes the complexity when measuring opinion. With more alternative options, one can gather more feedback from a user simultaneously. An interface is *pointwise* if it only measures the opinion regarding one item. Similarly, interfaces that measure information about two paired items are *pairwise*. With more than two options, the interface is *listwise*. For example, in Figure 5.1a, all interfaces are pointwise as they measure opinions regarding a single instance. Conceptually, a *listwise* interface also covers interfaces that potentially allow exploring infinite options (n=∞, e.g., Koyama et al. [83]). However, regardless of the involved number of options, since a user has to select or mark a subset of preferential relations among all options, e.g., Wilber et al. [157], a multiple-item selection can be considered a strict preference between selected and unselected ones, whereas the preference order among (un)selected ones remains unspecified. Note that although an interface presents paired options to measure human opinion, it could still ask humans to provide a distance estimation between two options, which only conveys absolute distance utility as a rating regarding one item to the reference. Furthermore, although Figure 5.2f presents four alternatives and asks how close they are to what the user is looking for, and this is an absolute rating only with respect to the user's idea of an optimum (which remains unknown to everybody else). These considerations imply that the underlying feedback type, which the framed UI presents as a usage instruction, is also crucial for opinion measurement.

The underlying type of feedback can be either *utility*-based or *preference*-based. Utility-based interfaces query a human to provide an absolute ordinal or cardinal value regarding an option (The interpretation of the value can differ depending on the context). Preference-based interfaces only measure relative information regarding paired options and, therefore, need further consideration to infer a latent utility. On the one hand, since utility-based interfaces only provide numeric information directly associated with the option, the relative information between options may not be explicit. However, the utility value can be sorted. On the other hand, preference-based interfaces only provide local information. A *hybrid* feedback type provides a subtle design consideration to mix utility- and preference-based information. Similarly, there are two design alternatives, either strict or weak, depending on the type of preference.

One can place the previously presented examples into this design space. Figure 5.1a contains 1-RS and n-RS: Twitter and Tinder only utilize binary responses regarding one option (like or not). The amazon by feature rating example is n-RS but evaluates multiple dimensions using scales. In Figure 5.1b, there are two examples of 2-AFC or 2-ANFC. Figure 5.1c shows three listwise interfaces: n-AFR (songs are ranked without allowing ties and without information about how they are being ranked), n-AFRD (options ranked with a distance, items with the same utility value are forced to be placed on different ranks).

Figure 5.2a and Figure 5.2b are n-ANFR because they require humans to rank options (permitting ties) without specifying a distance to other options, Figure 5.2c and Figure 5.2d are 2-AFC because they are judging two options and cannot tell indifference through the UI, Figure 5.2e is 2-ANFC because it offers “can’t tell” to its user, Figure 5.2f is n-RS, and Figure 5.2g is 1-RS because it is judging how far one item is away from the other, and only judges one option. Lastly, Figure 5.2h is n-AFR. In fact, Figure 5.2a and Figure 5.2h also use a meaningful 2D layout for presenting the choice alternatives instead of a plain list. While this may improve the efficiency and user experience in specific tasks, it does not change the general structure and, therefore, the classification of the interface.

As shown in Figure 5.3, one can select six generic interfaces from this design space (1-RS, 2-ANFC, n-RS, n-AFR, n-ANFR, n-ANFRD) to better facilitate the evaluation of human-in-the-loop optimization, see subsection 5.3.1. More specifically, regarding the exact behavior of these interfaces, 1) 1-RS uses a bipolar slider-based Likert scale, similar to n-RS; 2) n-AFR requires a full ranking of all options, and the interface does not allow expressing indifference in this case, whereas n-ANFR can, which characterizes strict and weak preferences; 3) To guarantee a pure ranking behavior in the interface, n-AFR, and n-ANFR automatically adjust the distance between options. For example, if a user dropped two boxes on the second and fourth choice container box. Then the fourth choice will automatically move up to the third choice and inform users that their ranking is automatically aligned and normalized. 4) n-ANFRD is the most flexible listwise ranking interface that permits users to place the option box in any given con-



UI1) 1-RS

UI2) 2-ANFC

UI3) n-RS

UI4) n-AFR

UI5) n-ANFR

UI6) n-ANFRD

Figure 5.3: Different opinion measurement interfaces: UI1) Single item Rating Scale (1-RS); UI2) Two Alternative Non-Forced Choice (2-ANFC); UI3) n-items Rating Scales (n-RS); UI4) n-Alternative Forced Ranking (n-AFR); UI5) n-Alternative Non-Forced Ranking (n-ANFR); UI6) n-Alternative Non-Forced Ranking with Distance (n-ANFRD). The 1-RS and n-RS interfaces are rating scale-based measurements, and the 2-ANFC, n-AFR, and n-ANFR are preference-based ranking interfaces. As a combination, the n-ANFRD mixes absolute ratings and preferential rankings, which permits expressing not only ranking orders but also local ordinal ranking distance. All interfaces in this gallery also permit users to express their incomplete preference through "I don't know" and to signal when a satisfactory result is achieved through "I'm satisfied".

tainer box. 5) All interfaces include either checkbox options or container boxes to allow users to express incomplete preferences.

5.3 Hypotheses and User Study

To answer our research questions regarding user interfaces in human-in-the-loop systems, we discussed our hypotheses and then conducted a user study to investigate how different opinion measurement interfaces impact the overall human-in-the-loop optimization performance.

5.3.1 Hypotheses

To compare the six classes of opinion measurement interfaces described above, this thesis formulates four major hypotheses concerning the performance of human-in-the-loop optimization when using such an interface. The details of how to measure performance were described in chapter 3.

Hypothesis 1: Baseline Early work on human-in-the-loop optimizations has shown great performance when using 2-ANFC [11, 12, 13]. Their results showed that 2-ANFC systems generally perform better than 1-RS and greatly support human-in-the-loop systems. It is also helpful to use their initial findings as the baseline comparison and to validate the experiment apparatus. Thus, it is reasonable to hypothesize (**H1**) that the *preference-based* interface will outperform the *utility-based* interface regarding performance measures (UI1 vs. UI2).

Hypothesis 2: Impact of Listwise Approach More, generally speaking, note that if **H1** is true, this would indicate that giving the user more options for comparison will improve the overall system performance. In fact, this is what findings by Cao et al. [18] suggest. Thus, one can hypothesize that a *listwise* comparison ($n > 2$), will outperform comparisons with fewer options. As *Listwise* is the most flexible in terms of feedback, the next investigation is reasonable to understand the different strategies that emerged from related work [82, 83, 100]. Hence, this thesis investigates design variations in which the human can either give preferential ratings (e.g., Koyama et al. [80]), provide a strict ranking selection (e.g., Koyama et al. [82]), a weak ranking selection (e.g., Marks et al. [96]), or a combination of those (hybrid). Consequently, one can hypothesize (**H2a**) that a *listwise* interface will outperform a *pointwise* interface when using *utility-based* feedback (UI1 vs. UI3). And one can hypothesize (**H2b**) that a *listwise* interface will outperform a *pairwise* interface when using *preference-based* feedback (UI2 vs. UI5).

Hypothesis 3: Impact of Listwise Design Variations Based on the initial findings by Mikkola et al. [100] on, one can additionally hypothesize **H3a** and **H3b**. In detail, one hypothesizes **H3a** that a *preference-based* listwise ranking interface will outperform a *utility-based* listwise rating interface (cf. [11, 12, 13]), while allowing for expressing indifference (UI3 vs. UI5). Further, regarding indifference, it is reasonable to hypothesize **H3b** that being able to express indifference in the interface (*weak preference*) will be better than a forced choice (*strict preference*) (UI4 vs. UI5).

Hypothesis 4: Impact of Hybrid Approach Based on hypotheses **H1**-**H3** and under the assumption that their outcome in this work is in line with prior work, this chapter proposes a new hybrid feedback interface. For this, ***n-ANFRD*** (UI6), as they consider best to support the human-in-the-loop optimization, thus, getting the most amount of information from the user. Thus, it is hypothesized that the hybrid listwise interface will outperform: **H4a** the pairwise interface (UI2), **H4b** the listwise rating interface (UI3), **H4c** the listwise strict ranking interface (UI4), and **H4d** the listwise weak ranking interface (UI5).

5.3.2 Participants

In order to validate the previously described hypothesis, an experiment that reflects the previously discussed study procedure and apparatus in chapter 4, was conducted to recruit participants worldwide on Prolific[2]. To guarantee high-quality responses, only consider participants must meet these requirements: 1) answered with consistent demographics, e.g., not more than five years of age difference in the study compared to the platform registration information, and 2) 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, longer than 1 second in each iteration to check the image and interact with the interface according to a pilot study observations. Therefore, the results are reported based on 360 participants (171 female, 185 male, and four diverse; age $\mu = 28.14$, $\sigma = 8.39$, range 18-66), with varied self-indicated English proficiency (CEFR scale[3]: A1 0.55%, A2 0.55%, B1 3.89%, B2 21.67%, C1 43.33%, C2 30.00%) and varied self-indicated photo color enhancement expertise (None 12.2%, Novice 46.7%, Intermediate 30.0%, Experienced 10.6%, Expert 0.6%). Each interface on each domain was tested by 30 participants.

5.3.3 Methodology

To analyze and discuss the collected data, we conducted a study on text summarization and photo color enhancement domains. For the usage of the interface, we measured participants' interaction details in each interface as *behavior measures*. Next, we recorded participants' provided feedback and corresponding system outcomes as *feedback measures* for the optimization loop. Lastly, we measured their subjective experience with the system after each completion as *questionnaire measures*. See chapter 3 for more details.

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.

[2] `https://prolific.co`, *last accessed 17.02.2023*

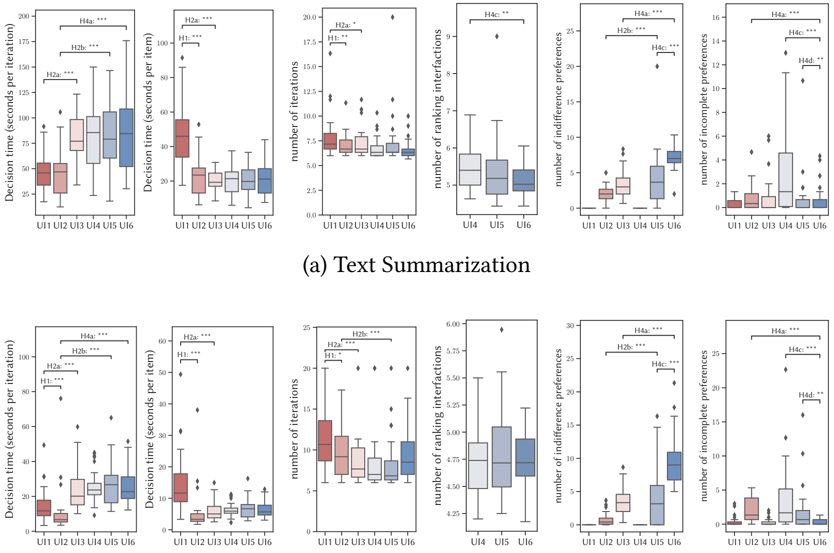
[3] `https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale`, *last accessed 17.02.2023*

In the subsequent analysis, we will present participants' behavior regarding interface interactions first, including their total iterations, decision time per iteration, and how they express indifference and incomplete preferences. Then, it shows the effectiveness of the feedback optimization loop concerning the objective quality and feedback utility. Lastly, subjective answers and comments are presented from a questionnaire. Since the hypotheses **H1** and **H2** in subsection 5.3.1 compare two UIs, one can analyze normally distributed data using a t-test or a Mann-Whitney-Wilcoxon test when normality is violated as indicated by the Shapiro-Wilk test [126]. Using linear mixed models [5, 88] (LMMs) is suitable to report the results for **H3**, **H4** consistently and statistically quantify the optimization loop. The models include participants as a random effect and compare pairwise and listwise UIs between groups. For non-significant independent two-sample tests, one can conduct either a Bayesian independent t-test or a Bayesian Mann-Whitney test and analyze the Bayes Factor [131, 152] to quantify the likelihood of hypotheses.

5.4 Behaviors and Interactions

Comparisons between all interfaces are shown in Figure 5.4 as visualization, and detailed statistics can be found in Ou et al. [113]. The following content will elaborate on the results regarding interaction behavior in all interfaces for both studied domains (text summarization and image color enhancement) in light of the hypotheses.

Decision Time For **H1**, for decision time per iteration, we found no significant difference between UI1 and UI2 in text summarization, but UI1 caused a significant increase in decision time compared to UI2 in image color enhancement. Regarding average decision time per item, there is a significant difference between UI1 and UI2. *This means that providing comparisons enables faster decisions.* For **H2**, UI3 results in a significantly longer decision time per iteration than UI1, and UI5 has a significantly longer decision time per iteration than UI2, which is true for both domains. This indicates that *users need more decision time to judge more given options.* However, when checking decision time per item, UI1 takes significantly more time than UI3 in both domains, while no significant difference was found between UI2 and UI5. This indicates that *the listwise approach did not increase decision time per item.* Finally, one can fit general linear mixed models for **H3** and **H4** using the Gamma family with an inverse link [94]. None showed significant differences in UI3-UI6 regarding decision time, except **H4a**: It compares UI2 to UI6, and the effect of UI2 is statistically significant and posi-



(a) Text Summarization

(b) Image Color Enhancement

Figure 5.4: Comparisons between opinion measurement interfaces, from left to right: **1)** *Decision time*: Listwise approaches (UI3, UI4, UI5, UI6) significantly increased decision time per iteration. However, decision time per item in listwise approaches does not differ from the pairwise (UI2); **2)** *Iterations*: Participants iterate significantly more in the pointwise (UI1) than in a pairwise (UI2) or listwise (UI3) approach; **3)** *Ranking interactions*: There is a significant difference between listwise interfaces, when indifference is not permitted (UI4 vs. UI6) in the text domain, but not in the image domain; **4)** *Indifference*: Participants express more indifference in listwise approaches if they can (UI5 and UI6) than in the pairwise approach (UI2.) Note that UI1 and UI4 cannot express indifference; **5)** *Incomplete preferences*: Participants express incomplete preferences significantly more when they cannot express indifference in listwise approaches (UI4 vs. UI5 and UI6.)

tive regarding decision time per iteration. *Although the pairwise interface enables faster total decision time when there are more options, it does not make per-item decisions significantly faster than the listwise approach.*

Interaction Behavior –Number of Iterations For the total iterations, participants explored significantly more with UI1 than with UI2 (**H1**) and UI3 (**H2a**) in both domains. However, there are no significant differences between UI2 and

UI3 (**H2b**). One can fit Poisson mixed models for **H3** and **H4**, and in total iterations, there were no significant differences for most of the comparisons between pairwise and listwise interfaces (UI2-UI6) in both domains. *While the pointwise interface (UI1) resulted in significantly more iterations than the other interfaces, there were no significant differences across domains between pairwise and listwise interfaces (UI2-UI6).*

Interaction Behavior –Number of Ranking Interactions From UI4 to UI6, listwise interfaces additionally recorded their ranking interaction for **H4**. In text summarization, a fitted Poisson mixed model showed that compared to the total ranking interactions in UI6 (β=1.627, SE=0.023, t=69.873, p<.001), the effect for UI4 is statistically significant and positive (β=0.065, SE=0.033, t=1.986, p=.047). For UI5, the effect is statistically non-significant and positive (β=0.037, SE=0.032, t=1.156, p=.248). However, in image color enhancement, compared to the total ranking interactions in UI6 (β=4.700, SE=0.064, t=73.761, p < .001), the effects for UI4 (β=-0.001, SE=0.092, t=-0.008, p=.994) and UI5 (β=-0.001, SE=.092, t=-.008, p=.479) are both statistically non-significant. In sum, *with the same BO optimizer under the hood, participants tend to produce fewer ranking interactions with UI6 than with other listwise interfaces (UI4 and UI5).*

Preference Type –Expressing Indifference Following the study design, UI1 and UI4 cannot express indifference for indifference preference. Thus, **H2a** and **H3b** are not evaluated here. For preference-based feedback, UI5 prompted participants to express significantly more indifference preference than UI2 (**H2b**). However, there is no significant difference between UI3 and UI5 for expressing indifference preference. For **H4a**, **H4b**, and **H4d**, one can fit a poisson mixed model. Compared to UI6. The effect of UI3 is statistically significant and negative compared to UI6. The effect of UI5 is statistically significant and negative. These results apply to both text summarization and photo color enhancement domains. Hence, *compared to all other UIs, participants express indifference significantly more often when using UI6.*

Preference Type –Expressing Incompleteness The Wilcoxon signed rank test shows participants who moved their sliders and used the "I don't know" checkbox in UI1 and UI4 because these interfaces cannot express indifference preference. There is no significant difference between adjusting the slider to the midpoint and clicking "I don't know" (W = 239.50, p = .219; r = -0.24, $CI_{95\%} = [-0.55, 0.13], BF_{10} = 0.638$). This result verifies the prior work regarding the indifference between incomplete preference and midpoint rating [2] in a bipolar Likert scale and hence validates the experiment apparatus.

In text summarization, for **H1**, participants with UI2 did not express significantly more incomplete preferences than with UI1, similar when comparing UI1 vs. UI3 (**H2a**, and UI2 vs. UI5 (**H2b**). For **H3**, a fitted Poisson mixed model shows, when compared to UI5, UI3 is statistically non-significant and positive, but the effect of UI4 is statistically significant and positive. Another fitted Poisson mixed model for **H4** shows that, compared to UI6, only UI4 is statistically significant and positive. Similarly, in image color enhancement, results showed that UI6 is expressed significantly less than UI2, UI4, and UI5. Therefore, *compared to all other UIs, participants tend to express fewer incomplete preferences when using UI6.*

5.5 The Optimization Loop

The overall optimization loop produces three types of data: 1) user feedback (absolute ratings or preferential rankings), 2) an optimizer-inferred (latent) utility function, and 3) the optimized system outcomes based on the learned utility function. As the objective of BO is to infer parameters that maximize the system outcome utility, if the optimization works for a metric, the system-proposed outcomes would come closer to user expectations over iterations; hence users would give higher utility responses. As an alternative verification of the optimization loop, analyzing the actual system outcomes using different objective quality metrics can help to understand how system outcomes change progressively. Detailed statistics are reported in Ou et al. [113].

Objective Quality of System Outcomes Since the user goal is to reduce the text length while preserving the meaning of the text or enhance the image color towards participants' preference, for the text domain, one can measure objective quality using ROUGE and BLEU, which compare the system outcome throughout the optimization loop. ROUGE measures how many human-summarized words appear in the machine-generated summaries, and BLEU measures precision regarding how many words in the machine-generated words appear in the human reference summaries. For the image domain, one can compute how five different metrics changed over iterations and look for consistent trends. Note the limitation of these metrics: they do not measure the semantic meaning of text summarization and only compare to one human writer's ground truth, provided in the CNN daily mail dataset. Similarly, the color metrics only correlate to human judgment and do not measure individual color preferences.

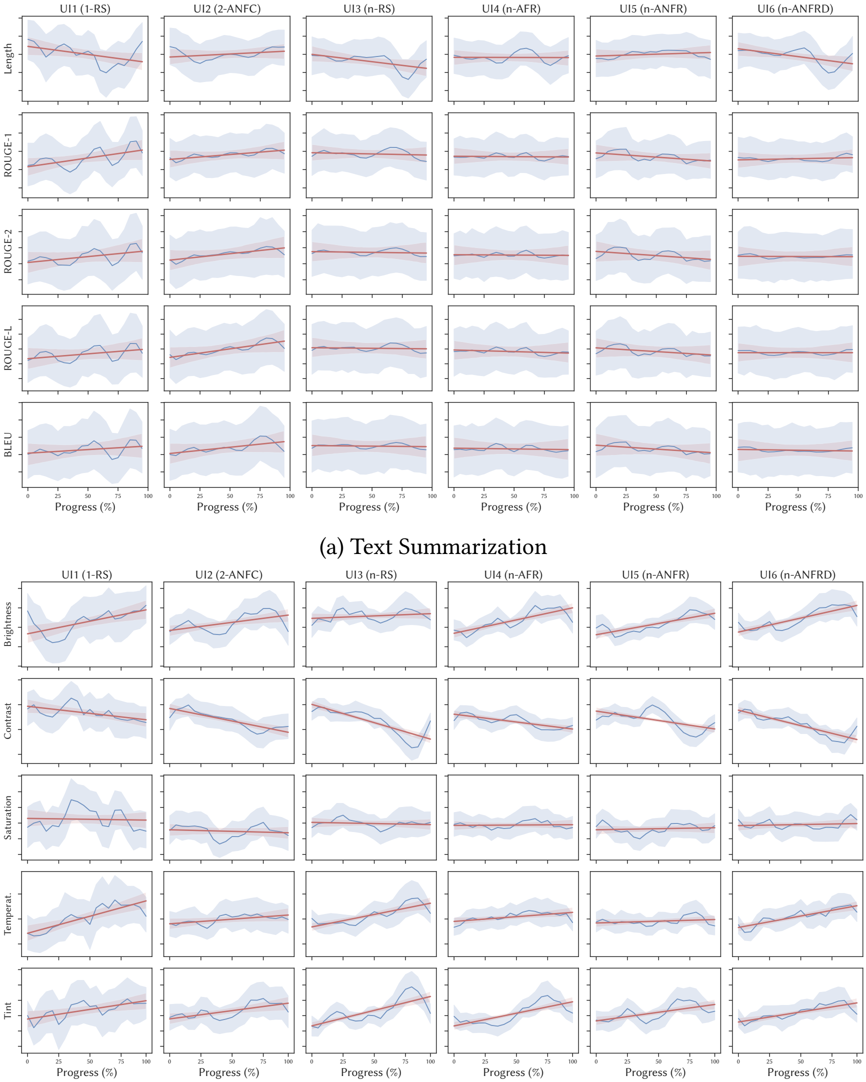
In summary, the results of the objective quality validate that *all interfaces can optimize objective metrics with a consistent trend.* In more detail, for the text summarization domain, the measured objective quality using *Length*, ROUGE

(*ROUGE-1*, *ROUGE-2*, and *ROUGE-L*) and *BLEU*, which compare the system outcome throughout the optimization loop. One can show the progress of the HITL optimization over time for the measurements in Figure 5.5a. Regarding summarized text *Length*, UI6 outperforms UI2 and UI4, and the effect of Progress is statistically significant. Moreover, UI6 outperforms UI3 with respect to *ROUGE-1*, and the effect of Progress is statistically significant for *ROUGE-1*. For the image enhancement domain, the thesis computed how five different metrics changed with Progress and looked for consistent trends. Namely *Brightness*, *Contrast*, *Saturation*, *Temperature*, and *Tint* are investigated, and the progress of the HITL optimization over time for the measurements is shown in Figure 5.5b. Here, UI6 outperforms UI3 in terms of *Brightness*. Moreover, the effect of Progress is statistically significant for all color metrics except *Saturation*.

User Feedback and Optimizer Learned Utility of Preferences When using UI1 and UI3, users were asked to provide absolute ratings of system-generated solutions (for utility-based optimizers), which can be consider as a *user feedback*. In contrast, in UI2, UI4, and UI5, users provide ranking responses (for preference-based optimizers) that give local preference relations between system outcomes. Based on this, the Bayesian optimizers then infer the *learned latent utility*. Lastly, in UI6, as it mixes absolute ratings and rankings, one can analyze both *user feedback* and the *learned latent utility*.

In total, four LMMs are fitted regarding *user feedback* and *lerned latent utility* independently from the Domain. We used Interface and Progress as fixed main effects to compare the performance between UIs; see Figure 5.6 and Figure 5.7. In terms of user feedback: For the text summarization domain, the LMM results showed the effect of iteration is statistically significant and positive, and UI6 outperforms UI1 and UI3. For the image color enhancement domain, the results show that UI6 significantly outperforms the other UIs. Regarding the *learned latent utility*: The results of the LMM for the text summarization domain showed the effect of iteration is statistically non-significant and negative. However, UI6 is significantly better than UI2, UI4, and UI5. For the image color enhancement domain, UI6 significantly outperforms the other UIs.

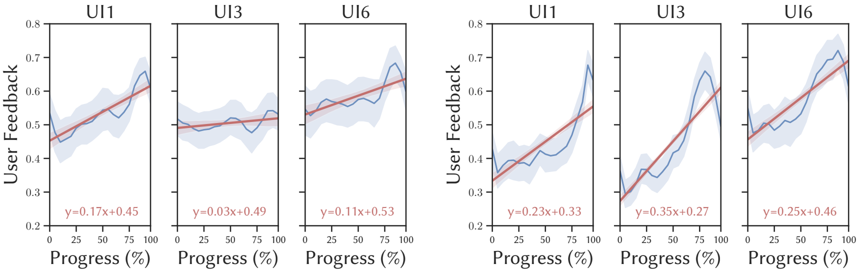
In summary, UI6 significantly outperforms the other UIs regarding optimizer *lerned latent utility* and *user feedback*. These results show that *through HITL optimizations, UI6 produced significantly higher utility than all other interfaces (UI1-UI5) in both domains.*



(a) Text Summarization

(b) Image Color Enhancement

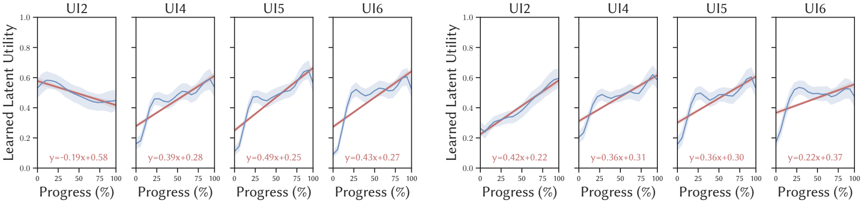
Figure 5.5: Comparing objective quality between interfaces. The blue line is the average utility. The red line is a linear regression of the blue line, with each individual slope indicated in red.



(a) Text Summarization

(b) Image Color Enhancement

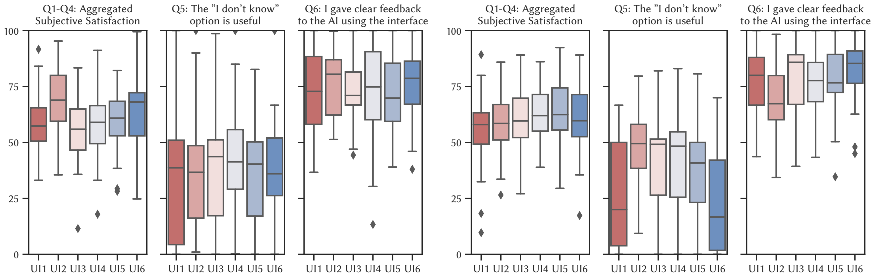
Figure 5.6: The *user feedback* (user input rating utility through the UIs directly) throughout the optimization loop. The blue line is the average *user feedback*, and the red line is a linear regression of the blue line. An increasing *user feedback* trends indicate the optimization goes more toward the users' preferences. Thus, increasing their feedback over time. Here, UI6 outperforms UI1 and UI3 in both domains after completing the optimization.



(a) Text Summarization

(b) Image Color Enhancement

Figure 5.7: The optimizer *learned latent utility* (inferred ranking utility through preferential Bayesian optimization) throughout the optimization loop. The blue line is the average utility, and the red line is a second-order polynomial regression of the blue line. An increasing utility (either direct or latent) trend indicates the optimized text is more toward users' preferences. In latent utility comparisons of the two domains, UI6 outperforms UI2, UI4, and UI5.



(a) Text Summarization (b) Image Color Enhancement

Figure 5.8: Questionnaire results were collected using a bipolar slider-based Likert scale from text summarization and image color enhancement tasks. 100 means "Strongly agree", and 0 means "Strongly disagree". Participants answered six questions per task. The results indicate significant differences between the pointwise (UI1) and pairwise (UI2) approaches, and the hybrid approach (UI6) outperforms other listwise approaches (UI3, UI4, and UI5) but produced no significant differences compared to the pairwise interface (UI2). Participants considered expressing incomplete preference using "I don't know" as less useful, and they all thought they gave clear feedback.

5.6 Questionnaire Results

As an additional side verification, participants are asked to answer six questions as shown in Figure 5.8.

In text summarization, for **H1**, participants did not report significantly higher satisfaction in UI2 (M=63.75, SD=18.66) compared to UI1 (M=57.02, SD=20.32). For **H2a**, also UI1 compared to UI3(M=59.01, SD=18.45) showed no significant difference. Finally, for **H2b**, the results show a significant difference when comparing UI2 to UI5 (M=73.67, SD=20.39). For **H3** and **H4**, there are no significant differences between all interfaces (UI2-UI6) in terms of satisfaction. Similarly, allowing expressing incompleteness preference in UI6 is significantly less helpful compared to UI2-UI5 (all p<.050) but not when compared to UI1 in the image color enhancement domain. For Q6, participants felt they gave more clear feedback when using UI6 compared to UI2 ($\beta = -13.933$, $SE = 3.851$, $t = -3.618$, $p < .001$) - **H4a** in image color enhancement domain.

These questionnaire results show that participants were significantly more satisfied (Q1-Q4) with listwise interfaces than with a pairwise interface. They reported that

UI6 was clearer in expressing their preference and hence expressed fewer incomplete preferences (Q5) and could give more clear instructions (Q6).

5.7 Discussion

Overall, a significant difference was found in all other interfaces to outperform the pointwise interfaces in both studied domains, but there are subtle differences between listwise and pairwise UIs. In this section, more structured results will be discussed according to the initial hypotheses and summarize possible implications and tradeoffs when determining the use of an interface for human-in-the-loop optimization.

5.7.1 Validating Hypotheses

H1: Baseline Apparatus Verification Based on the results, the 2-ANFC pairwise interface (UI2) outperforms the 1-RS pointwise interface (UI1) because it has a significantly lower decision time per item and fewer iterations. The 2-ANFC also helped participants to express indifference and their incomplete preference for better system parameter exploration and exploitation. Subjects using 2-ANFC also reported significantly higher satisfaction compared to 1-RS. This result verifies that the experimental apparatus is valid and aligns with discussions in prior work [11, 12, 13].

H2: The Listwise Approach In terms of UI interactions, the results showed that listwise interfaces (n-RS/n-ANFR, UI3/UI5) provide the following advantages (compared to 1-RS/2-ANFC, UI1/UI2): 1) they allow users to evaluate more options without increasing decision time and iterations; 2) users express more indifferent preferences and less incomplete preferences; 3) they allow implicit comparisons among all given options. However, the direct use of the listwise approach (i.e., using more sliders to rate more items than 1-RS or rank multiple items without distance in n-ANFR) did not significantly improve the performance of the optimization loop.

H3: Design Variations of The Listwise Approach When comparing design variations regarding querying different feedback types for ranking (**H3a**: n-RS vs. n-ANFR, i.e., UI3 vs. UI5), allowing expressing indifference and incomplete preferences (**H3b**: n-AFR vs. n-ANFR, i.e., UI4 vs. UI5) in listwise approach, there is no enough evidence to show a difference in decision time, iterations, and the performance of the optimization loop. The subjectively reported satisfaction also did not differ significantly from each other. However, n-AFR expresses significantly more incomplete preferences compared to n-ANFR.

H4: A Hybrid Approach The results showed that the hybrid approach (n-ANFRD, UI6) did not differ from 2-ANFC (**H4a**), n-RS (**H4b**), n-AFR (**H4c**), n-ANFR (**H4d**) in terms of decision time per item and involved iterations, and subjective satisfaction. However, the suggested interface let participants express indifference significantly more than other interfaces, resulting in fewer ranking interactions. When checking the performance of the optimization loop, n-ANFRD showed significantly higher directly measured utility throughout the loop and higher latent preferent utility compared to pairwise and other listwise interfaces.

5.7.2 Interpretation and Implications

While the experiment initially assumed using the listwise approach that increased the number of parallel comparisons could improve the overall performance of the optimization loop, the results did not support this assumption. Based on the validation of **H3**, neither n-RS (UI3, utility-based) nor n-ANFR (UI5, preference-based) show a significant advantage and only increased the overall decision time in the loop compared to 2-ANFC (UI2, preference-based). This might be because neither n-RS nor n-ANFR provided enough support for participants to reduce the overall noisy observations, and there is room for improvement when using a BO optimizer specifically designed for listwise UIs. In n-RS, despite multiple given alternatives, participants might not notice the relative position between different sliders; in n-ANFR, the UI did not demonstrate enough absolute cues to participants on how pure ranking can influence their input, and a pure local ranking may not give participants information on a global overview.

Instead, the n-ANFRD hybrids collect utility-based and preference-based feedback by providing rating scales in the regular listwise ranking UI. Similar to other listwise UIs, based on the validation of **H4**, although n-ANFRD increased total decision time, the results show the general advantage of preference elicitation (expressing more indifference, fewer ranking interactions, fewer incomplete preferences, and better performance in the optimization loop). This could be interpreted as the UI starting to offer implicit recall support for participants to their past ratings if their ranking positions are improving. In addition, n-ANFRD resulted in fewer ranking interactions and made it clearer for participants to express a ranking decision compared to other purely listwise ranking interfaces.

Furthermore, the results regarding designing UIs to support indifference and incomplete preferences offered implications for future modeling algorithms. In Bayesian optimization, indifference could be an indicator to inform the acquisition function to avoid exploiting these areas; incomplete preferences might be another indicator to balance exploration and exploitation. Depending on the context, one might design the algorithm to a) avoid exploring a space where people had an incomplete preference to reduce input uncertainty, or alternatively b) explore it even more for creativity support and then integrate the measured preferences with more interaction behavior, such as interface features and overall task completion time [32], to better support the optimization process.

5.7.3 Design Trade-offs

As validated in the hypotheses, there are many subtle differences between pairwise and listwise interfaces when permitting users to express indifference and incomplete preferences. This gives us evidence of how they could impact human-in-the-loop optimization performance. The following content discusses potential tradeoffs when using these interfaces.

Guideline 1: Pointwise Approach The pointwise slider-based interface has proven useful in many disciplines, such as psychology, subjective scale measures, etc. It is useful to identify social agreement among populations and measure the direct utility given by a human. The technique for processing these types of data is also mature. But it is very unstable and problematic when directly applying it to individual sequential measurements. Future human-in-the-loop optimization systems or any sequential measurement should generally avoid using this approach.

Guideline 2: Pairwise Approach A pairwise interface only requires a human to judge based on comparing two given items. Since there is no measured utility and it only gives relative information about two items, the modeling of the latent utility function could suffer from a change of criteria and intransitive preference when the anchoring point between two sessions is not chosen properly. When an iteration receives an incomplete preference, the underlying optimizer cannot receive much useful modeling information. However, when a design objective is to optimize for fast decisions, the pairwise approach (2-ANFC) could be a winner among other alternative listwise designs.

Guideline 3: Listwise Approach Listwise interfaces establish a comparison by increasing the number of alternatives. Similar to the pairwise interface, a listwise interface without measuring ranking distance (n-AFR and n-ANFR in this thesis) requires a properly sampled anchoring option between iterations to establish a ranking function. Otherwise, there is no direct association between iterations. Instead, a listwise interface that allows expressing ranking distance (n-RS and n-ANFRD in this thesis) conveys an implicit association between different iterations regardless of whether the human provides local or global judgments it is more suited to use in a human-in-the-loop optimization context. Furthermore, n-ANFRD is better than n-RS because a concentrated participant may not notice the relative slider position between parallel judging options in n-RS, but n-ANFRD explicitly visualizes the ranking in the UI for users, which helps users more clearly express their opinion in a sequential task. As design suggestions, when there is a limited measuring budget, e.g., one can only query humans for a very limited number of opinions, applying listwise interfaces is considered better than pairwise because it provides more measurement of their opinion and does not decrease the optimization performance. Furthermore, enabling users to express ranking distance (n-ANFRD) is the winner of all listwise variations (n-RS, n-AFR, and n-ANFR).

Guideline 4: Measuring Indifference and Incomplete Preferences When measuring more accurate user opinions, allowing them to express indifference and incomplete preferences is crucial. For the pointwise interface, these two different types of preference were implicitly ignored and treated as being indicated by centering the slider. This interpretation of a central slider position may be true with proper task wording. Still, it may be easily misused and produce more noisy input when a user does not know how to express their opinion with the interface. Therefore, in the human-in-the-loop optimization scenario, it is better to explicitly design these two functionalities to support users in expressing their indifference and incomplete preferences. This information could also inform optimizers to avoid exploring these areas. This is particularly important when the evaluation budget is very limited in the human-in-the-loop optimization context.

5.7.4 Limitations

This chapter examined a text summarization task and an image color enhancement task to explore the impact of interfaces in human-in-the-loop optimization. While these domains are common and make it easy to run a study online at scale, and the results across these two domains are consistent with each other, it is not verified in more domains.

Since the optimization context tightly relies on humans and modeled human feedback as objective functions, the system performance highly depends on the quality of human input. The responses collected online may be influenced by participants' environments, such as slow decision time due to external distractions. Moreover, for the preferential settings, depending on the underlying optimizer, the decision time might be different if the system repeatedly returns the current optimum (e.g., [12]).

The experiment in this chapter evaluated opinion measurement interfaces, not optimizers. However, differently designed listwise interfaces could benefit from dedicated Bayesian optimizers. Without a specialized BO optimizer, the suggested n-ANFRD interface supports the user to express a ranking distance and still outperforms other interfaces in various aspects, such as fewer ranking interactions, more clearly stated indifference, and rarely expressed incomplete preferences. This chapter also did not look into a detailed inspection of the cause of indifference and incomplete preference.

Lastly, the discussion of this chapter focused on the design space for aggregated preferences and used Bayesian optimizers that assumed one latent preference function. However, the latent preference function could be limited by the human preference aggregation process, where multiple criteria may not be expressed simultaneously.

5.8 Summary

In summary, with an initial discussion of multiple opinion measurement interfaces, an experiment and corresponding analysis of the results showed the impact of interface design on human-in-the-loop optimization performance. Both n-ANFRD and 2-ANFC have their unique advantages, and the design decision between them is a tradeoff of decision time, the precision of measured human feedback, and overall optimization loop performance.

In the following chapters, this thesis will use n-ANFRD to narrow down the solution space of design principles and continue to discuss other building blocks in human-in-the-loop systems.

6 Termination Condition

No other question has ever moved so profoundly the spirit of man; no other idea has so fruitfully stimulated his intellect; yet no other concept stands in greater need of clarification than that of the infinite.

– David Hilbert

When a human user starts to interact with an intelligent machine, the observable behaviors between the user and the machine UI are often intertwined in a complex way. Based on a higher-level analysis of the interaction-optimization loop, chapter 5 has explored the most effective user interface to measure a user's opinion regarding the machine's proposed outcomes. However, the study presented in chapter 5 configured a limited number of iterations and terminated automatically when it exceeded the maximum allowed iterations. When will the interaction loop terminate? Will the user always terminate the interaction proactively? In this chapter[1], a domain-specific example in 3D model processing showcases and further discussion from the user's perspective to reveal how human-in-the-loop optimization systems can fail unexpectedly without carefully considering implicit system assumptions. The overall results imply two different types of violations of underlying system assumptions: 1) the human's judgment is highly heuristic and inconsistent in a sequential decision-making context; 2) the ma-

[1] The content of the chapter is partly based on Ou et al. [112].

chine algorithms contain assumptions that are context-specific and fixed, which may violate the applied reality.

6.1 Hypotheses and User Studies

To evaluate the effectiveness of human-in-the-loop systems from a user perspective, we present two studies in this chapter; first, a field study with an industrial partner, and a lab study which verifies the effects of the field study. In the field study, we used a 3D model processing system in a real-world setup with our industrial partner, who aimed to optimize their process in customer projects. Here, we ran 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.

6.1.1 Field Study

We first conducted small pilot experiments for the field study to fine-tune our system's parameters to fit the partner's needs and customer projects.

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 [45, 100] the system to converge to the desired outcome quickly.

The experts used our system almost daily to evaluate model quality during polygon reduction. However, they were not restricted to our interface. They could use further software aids (as in their previous workflow) for the model quality inspection, e.g., for accessing more professional curvature visualizations. When

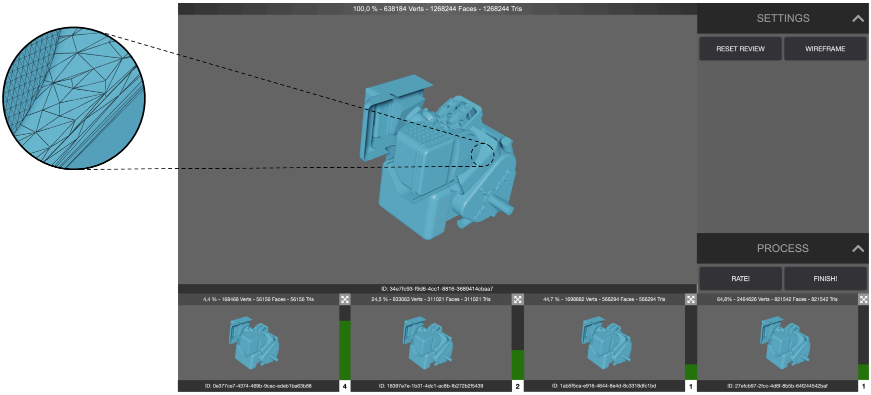

100,0 % - 638184 Verts - 1268244 Faces - 1268244 Tris
SETTINGS
RESET REVIEW
WIREFRAME
PROCESS
RATE!
FINISH!
ID: 34e7fc93-f9d6-4cc1-8816-3689414cbaa7
4,4 % - 168468 Verts - 56156 Faces - 56156 Tris
24,5 % - 933063 Verts - 311021 Faces - 311021 Tris
44,7 % - 1698882 Verts - 566294 Faces - 566294 Tris
64,8% - 2464626 Verts - 821542 Faces - 821542 Tris
ID: 0e377ce7-4374-469b-9cac-edeb1ba63b88
4
ID: 18397e7e-1b31-4dc1-ac8b-fb272b2f5439
2
ID: 1ab5f5ca-e916-4644-8e4d-8c3318dfc1bd
1
ID: 27efcb97-2fcc-4d6f-8b5b-64f244542baf
1

Figure 6.1: 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 [8] because 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).

using our tool, the loaded 3D model was computed into four variants. After the experts finished their evaluation (either inside or externally), they rated the models in our interface. These ratings were used for the next optimization iteration to generate new variants (see Figure 6.2). The rating process terminated when the experts found the results satisfactory or reset it.

6.1.2 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.

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 Figure 4.1. In detail, we asked participants to balance the trade-off between polygon reduction and quality loss. Thus, they had to indicate their preference

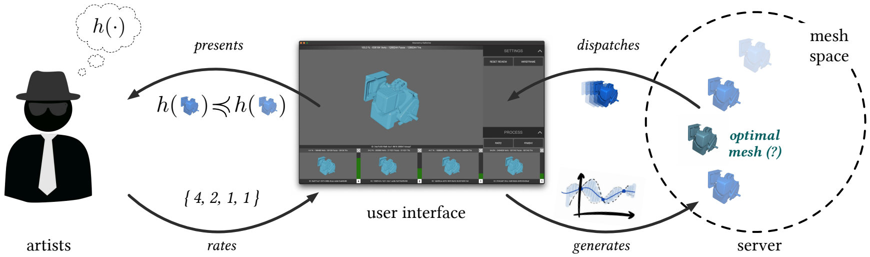


Figure 6.2: A human-in-the-loop 3D model processing system: A server generates differently processed variations of a complex 3D model and dispatches them to a user interface, which presents those variants to a 3D artist, who in turn rates them. Based on these ratings, new parameter settings are generated, 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.

using ratings to optimize models iteratively.

We selected five different 3D models, as shown in Figure 6.3, and each model[2] 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.

6.1.3 Participants

In the field study, 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.

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

[2] 3D mesh artifacts are provided courtesy of WAY Digital Solutions, Jeff H, Jose Olmedo, kenik, yarulesemel, and Stephan Thieme.

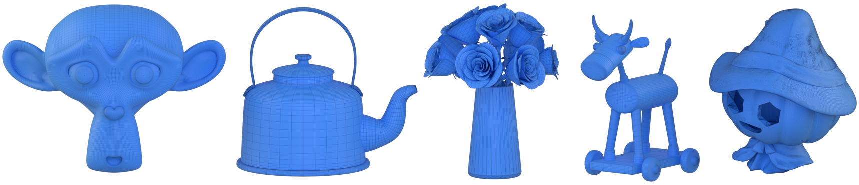


Figure 6.3: Models that are selected in the lab study, from left to right: *monkey, teapot, rose, cow, pumpkin.* Participants ranked each model's variants in each iteration regarding the quality of mesh simplification. Each model was presented to the participant twice (with/without a wireframe).

optimization requested. The remaining 134 sequences (number of iterations: $\mu = 4.1, \sigma = 4.2$, range 1-23) contain sequential preference ratings. The 3D models included various meshs, including organic, soft surfaces, hard technical surfaces with sharper angles, and combinations, such as machinery parts (see Figure 6.1) with both smooth, flowing lines and hard, mechanical edges.

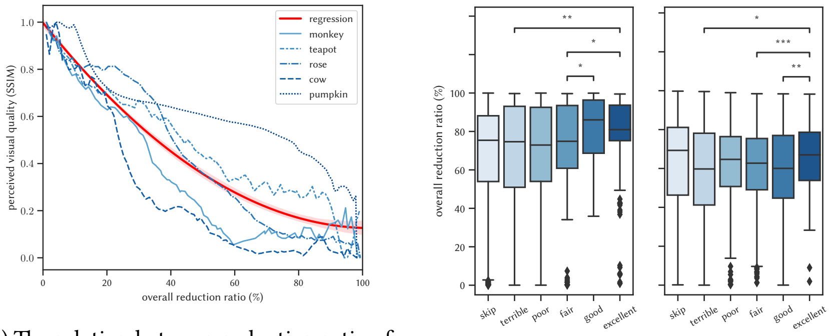
In the lab study, we recruited 20 participants using convenience sampling (7 female and 13 male; age $\mu = 27.0, \sigma = 8.8$, range 18-62). Four had more than a year of industrial experience in 3D modeling, and all others had no experience.

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.

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.

6.2 Human-AI Mutual Interventions

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*,



(a) The relation between reduction ratio of models in lab study and human perceived visual quality (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 6.4: Overview of the visual influences of polygon reduction and collected rating data.

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 human-in-the-loop outcomes.

Effectiveness of Human Judgments Figure 6.4a shows a second-order polynomial regression between the 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) [153] 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 6.4b shows the rating distributions in the two studies regarding the reduction ratio.

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

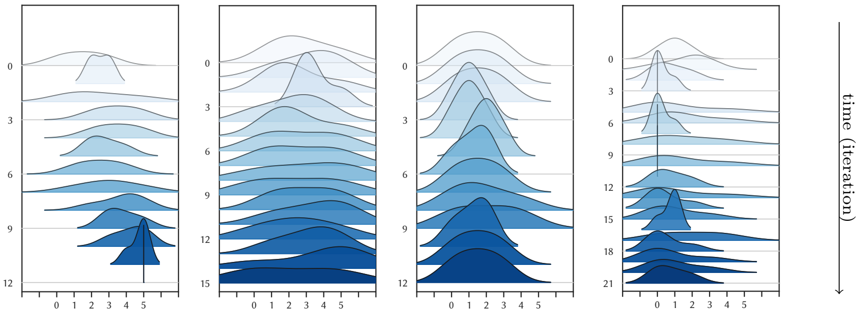


Figure 6.5: Comparing an ideal (far left) and actual (the others) rating distributions over time (from top to bottom), the bottom axis 0 to 5 represents the rating scale. An expected rating distribution should move to the right side over time if users are more satisfied with the results, but the actual preferences drift back and forth between 0 (*skip*) and 5 (*excellent*).

found a significant difference between *fair* (M=74.88) and *excellent* (M=80.95) ratings (U=8756.0, $p < .001$) and a significant difference between *terrible* (M=74.65) and *excellent* (M=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* (M=74.65) and *poor* (M=72.93) ratings (U=19705.0, $p = .62$) or *good* (M=80.03) and *excellent* (M=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* (M=60.03) and *excellent* (M=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 6.4b), which suggests that *lab study participants are easier to satisfy by the system outcomes than expert artists.*

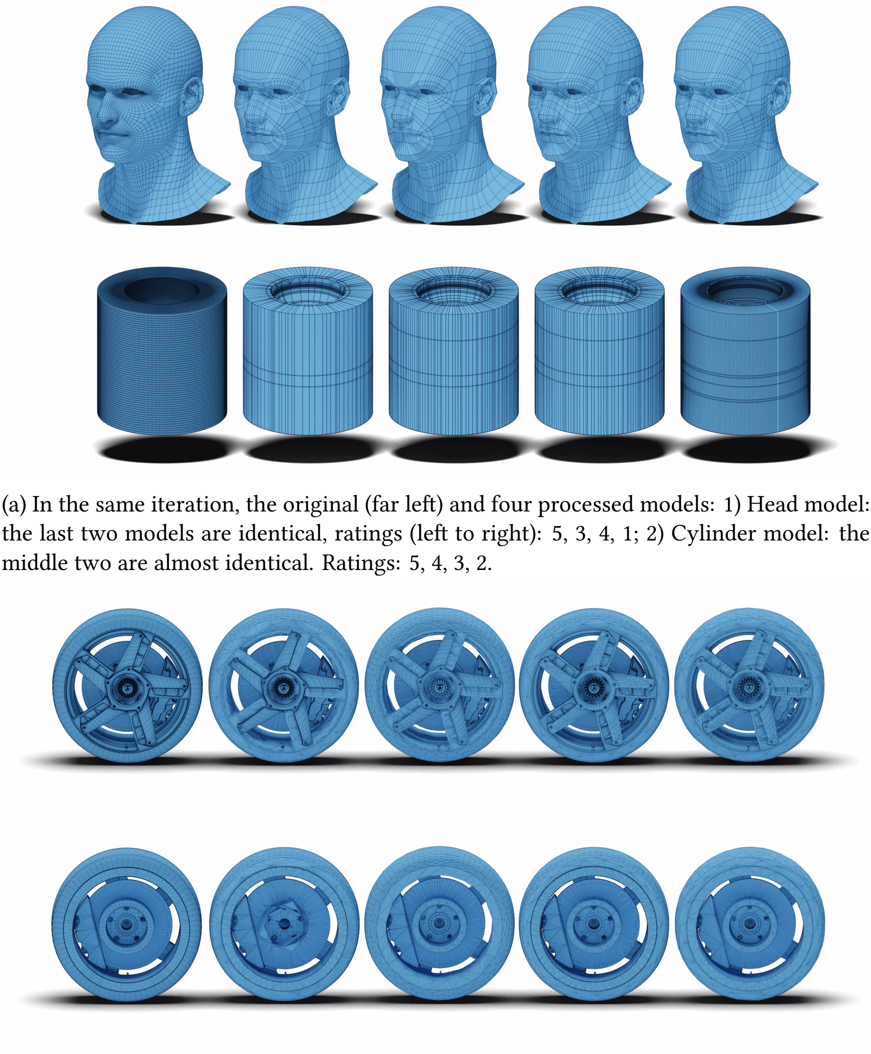
Stationarity and Trends of Data Figure 6.5 compares how an ideal (far left) and three actual rating distributions (the rest) drift over time: In our context, since the objective of using PBO is to search a polygon reduction configuration to maximize the human ratings [45], ideally, in a successful exploring 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 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.*

6.3 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 6.6 shows three of the discussed models. Figure 6.6a 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 detail: Ironically, the two head models on the right are identical but received an entirely different rating score of 4 and 1 from the same rater in the same round (*case 1*). The right middle 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 6.6b), 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 some-*



(a) In the same iteration, the original (far left) and four processed models: 1) Head model: the last two 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 four 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 6.6: Selected examples that were discussed in the semi-structured interviews.

times 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 right middle model has an objectively higher reduction ratio, and both contain 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 five 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.*"

6.4 Discussion and Implications

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 [125, 137]). Since errors are further propagated and amplified to system outcomes, we combine theories regarding human decision errors to reflect on and explain our findings.

6.4.1 Errors from Human

Based on our observed instability and expert feedback, we argue that 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 6.3), 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. Case 3 matches this bias as the artist decides based on his professional experience. c) *Representativeness* shows that decisions made by substitution examples may

occasionally be biased. Case 2 shows this behavior because the actual decision used a 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 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 human-in-the-loop 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.

6.4.2 Errors from Machine

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 human-in-the-loop 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 pre-

requisites 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 human-in-the-loop 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 is 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.

6.4.3 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 human-in-the-loop 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:

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;

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.

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.

6.4.4 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 through 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 [67] 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.

6.5 Summary

In this chapter, we discussed a human-in-the-loop system where an optimization algorithm in the background exploits sequential user choices to adapt 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 collected user interaction data and interview discussions with expert artists, we reflected on concrete influences that can break preference-optimized human-in-the-loop 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 suggest 1) 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 that human-in-the-loop may not always meet the termination criteria. Optimize polygon reduction tasks using the human-in-the-loop strategy requires resolving these issues; 2) the observed constraints also apply in a similar human-in-the-loop optimization context, and we proposed descriptive UI design directions as promising countermeasures to prevent human-in-the-loop optimization system outcomes from being highly unstable and eventually non-satisfactory.

7
Expertise and Objective Alignment

Anything that gives us new knowledge gives us an opportunity to be more rational.

– Herbert Simon

Suppose a human-in-the-loop system is bound to fail when the underlying machine algorithm assumption is strongly challenged. How can we improve the overall situation and reach a successful outcome? How does user expertise impact a human-in-the-loop system's overall exploration and exploitation process?

This chapter[1] examine the objective alignment process between user and system, and explore the relationship between human expertise and subjective satisfaction regarding system outcomes in text, photo, and 3D mesh optimization contexts. The results show 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. These results mean that experts tend to seek more diverse outcomes. At the same time, 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. These findings not only contradict the intuition that higher expertise will lead to better

[1] The content of the chapter is partly based on [114].

results but also informs future research, and practitioners should be cautious in dealing with the influence of user expertise when designing human-in-the-loop systems.

7.1 Hypotheses and User Study

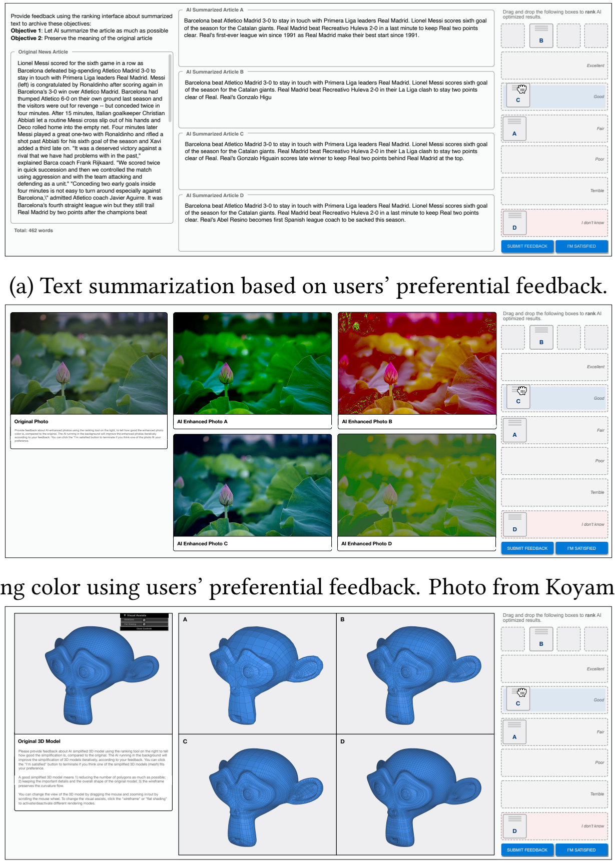
To understand the impact of expertise on satisfaction, one can hypothesize that by using human-in-the-loop optimization, participants with a higher level of expertise will produce a better outcome quality and perceive higher satisfaction than novice participants. To verify this, this chapter designed a between-group controlled experiment in three problem contexts: text summarization, photo color enhancement, and 3D model simplification. As dependent variables, the experiment measured participants' expertise in a domain context, interactions with the system, and feedback from final questionnaires (individual rating scales and open questions).

Figure 7.1 show our UIs in the human-in-the-loop 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.

With a detailed exploration of the interface design in chapter 5, 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 data, 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.

7.1.1 Participants

Similar to chapter 5, the overall experiment procedure in this chapter was shown in Figure 4.1. The involved participants were recruited worldwide on Prolific.



Provide feedback using the ranking interface about summarized text to archive these objectives:
Objective 1: Let AI summarize the article as much as possible
Objective 2: Preserve the meaning of the original article

Original News Article

Lionel Messi scored for the sixth game in a row as Barcelona defeated big-spending Atletico Madrid 3-0 to stay in touch with Primera Liga leaders Real Madrid. Messi (left) is congratulated by Ronaldinho after scoring again in Barcelona's 3-0 win over Atletico Madrid. Barcelona had thumped Atletico 6-0 on their own ground last season and the visitors were out for revenge -- but conceded twice in four minutes. After 15 minutes, Italian goalkeeper Christian Abbiati let a routine Messi cross slip out of his hands and Deco rolled home into the empty net. Four minutes later Messi played a great one-two with Ronaldinho and rifled a shot past Abbiati for his sixth goal of the season and Xavi added a third late on. "It was a deserved victory against a rival that we have had problems with in the past," explained Barca coach Frank Rijkaard. "We scored twice in quick succession and then we controlled the match using aggression and with the team attacking and defending as a unit." "Conceding two early goals inside four minutes is not easy to turn around especially against Barcelona,\" admitted Atletico coach Javier Aguirre. It was Barcelona's fourth straight league win but they still trail Real Madrid by two points after the champions beat

Total: 462 words

AI Summarized Article A

Barcelona beat Atletico Madrid 3-0 to stay in touch with Primera Liga leaders Real Madrid. Lionel Messi scores sixth goal of the season for the Catalan giants. Real Madrid beat Recreativo Huleva 2-0 in a last minute to keep Real two points clear. Real's first-ever league win since 1991 as Real Madrid make their best start since 1991.

AI Summarized Article B

Barcelona beat Atletico Madrid 3-0 to stay in touch with Primera Liga leaders Real Madrid. Lionel Messi scores sixth goal of the season for the Catalan giants. Real Madrid beat Recreativo Huleva 2-0 in their La Liga clash to stay two points clear of Real. Real's Gonzalo Higu

AI Summarized Article C

Barcelona beat Atletico Madrid 3-0 to stay in touch with Primera Liga leaders Real Madrid. Lionel Messi scores sixth goal of the season for the Catalan giants. Real Madrid beat Recreativo Huleva 2-0 in their La Liga clash to stay two points clear of Real. Real's Gonzalo Higuain scores late winner to keep Real two points behind Real Madrid at the top.

AI Summarized Article D

Barcelona beat Atletico Madrid 3-0 to stay in touch with Primera Liga leaders Real Madrid. Lionel Messi scores sixth goal of the season for the Catalan giants. Real Madrid beat Recreativo Huleva 2-0 in a last minute to keep Real two points clear. Real's Abel Resino becomes first Spanish league coach to be sacked this season.

Drag and drop the following boxes to **rank** AI optimized results.

(a) Text summarization based on users' preferential feedback.

Original Photo

Provide feedback about AI-enhanced photos using the ranking tool on the right, to tell how good the enhanced photo color is, compared to the original. The AI running in the background will improve the enhanced photos iteratively according to your feedback. You can click the "I'm satisfied button to terminate if you think one of the photo fit your preference.

Drag and drop the following boxes to **rank** AI optimized results.

(b) Enhancing color using users' preferential feedback. Photo from Koyama et al. [82].

Original 3D Model

Please provide feedback about AI simplified 3D model using the ranking tool on the right to tell how good the simplification is, compared to the original. The AI running in the background will improve the simplification of 3D models iteratively, according to your feedback. You can click the "I'm satisfied" button to terminate if you think one of the simplified 3D models (mesh) fits your preference.

A good simplified 3D model means 1) reducing the number of polygons as much as possible; 2) keeping the important details and the overall shape of the original model; 3) the wireframe preserves the curvature flow.

You can change the view of the 3D model by dragging the mouse and zooming in/out by scrolling the mouse wheel. To change the visual assists, click the "wireframe" or "flat shading" to activate/deactivate different rendering modes.

Drag and drop the following boxes to **rank** AI optimized results.

(c) 3D mesh simplification based on users' preferential feedback. The 3D model is a standard Blender Suzanne model. To inspect the rendered 3D models, participants can zoom in/out, pan, and rotate all models simultaneously.

Figure 7.1: The ranking interface for a) text summarization, b) photo color enhancement, and c) 3D model simplification. 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 or "skip" the entire ranking iteration.

Because participants had different median completion times in different experimental conditions, they are paid between £3 to £9 upon completion, corresponding to an hourly wage of €10.37/h ($11.47/h). Participants gave informed consent at the beginning of the study; thus, the study adhered to European privacy laws (GDPR). In total, the collected data were taken from 91 participants from 13 countries.

To guarantee high-quality responses, these criteria are considered: 1) a participant has an approval rate of 95%, 2) a participant completed the study only once, 3) a participant 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) a participant 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.

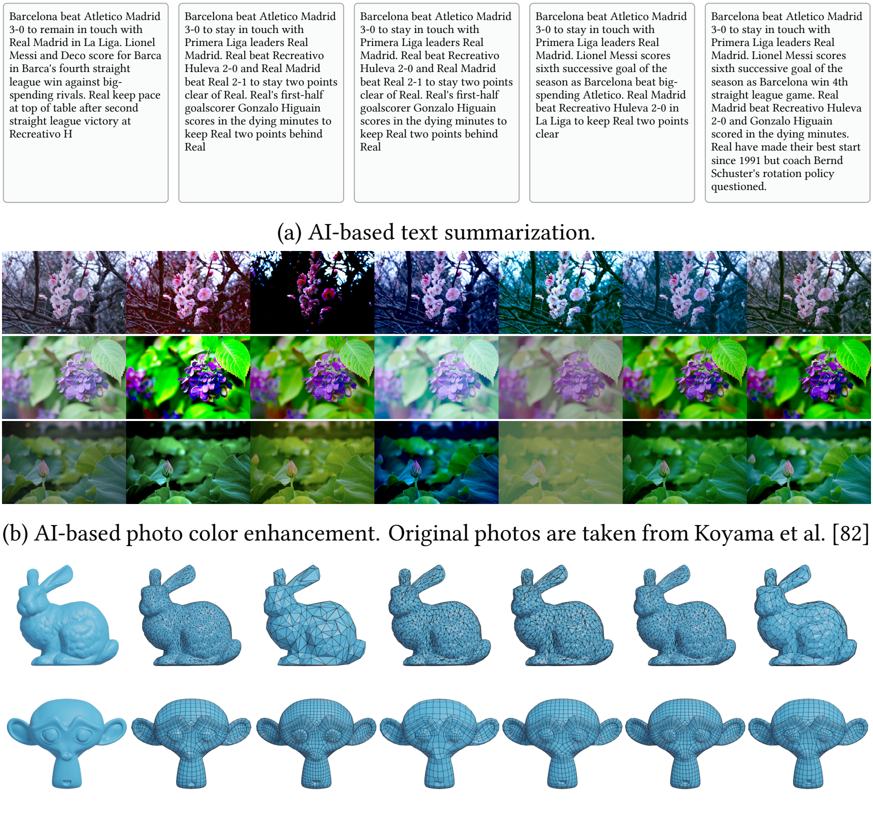
Therefore, this chapter will report results based on 60 participants (31 female, 29 male, and no diverse; age $\mu = 26.92, \sigma = 6.44$, range 19-52). Each domain context includes 20 participants. Example iteratively optimized outcomes are shown in Figure 7.2.

7.1.2 Inferring Levels of Expertise

Our participants reported varied experiences in different domains. They self-indicated English proficiency on the CEFR scale[2]: 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 five years 20%; for photo editing: No work experience 50%, less than one year of experience 10%, 1 to 5 years 30%, more than five 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 two weeks 20.0%, two 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 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%,

[2] `https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale`, *last accessed 17.02.2023*



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. [82]

(c) AI-based 3D model simplification.

Figure 7.2: 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).

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).

7.2 Behavior and Subjective Satisfaction

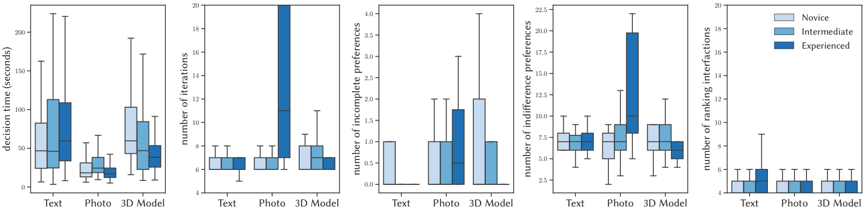


Figure 7.3: Measured interactions of participants in different domain contexts. Measurements are grouped by the level of expertise. The results indicate that experienced participants express their preferential ranking decisions more clearly than novices. For example, they behave faster in decision time with more iterations or decide slower with more ranking interactions (thoughtful indecision); they also express fewer incomplete and more indifferent preferences.

7.2.1 Interaction Behaviors

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.

All measured interaction behavior indicators are visualized in Figure 7.3. 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 significantly 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.*

7.2.2 Subjective Satisfaction

As mentioned in section 3.2.1, 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 7.4.

We conducted an ART ANOVA [160], 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).

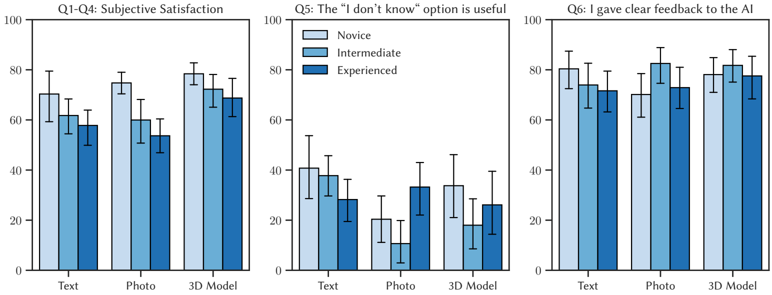
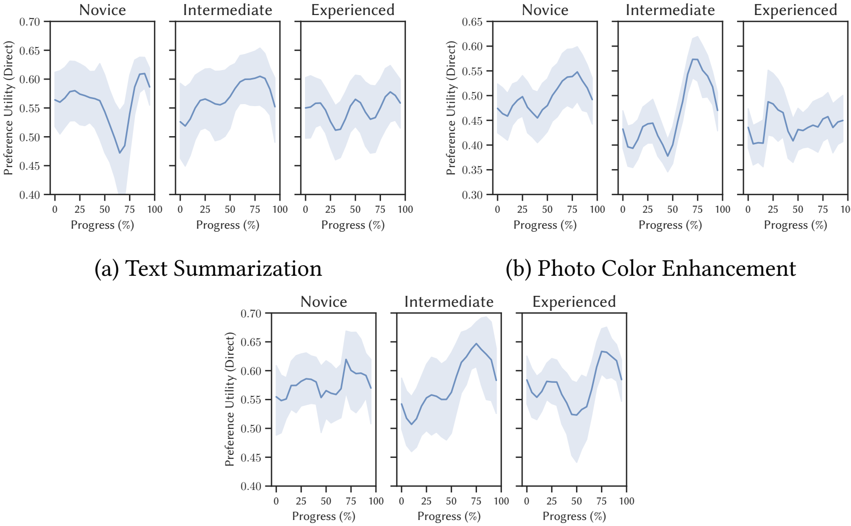


Figure 7.4: Measured subjective satisfaction, the usefulness of providing incomplete preference option while doing the ranking evaluation. The results suggest that subjective satisfaction significantly decreases when comparing novice and experienced participants. All participants considered allowing expressing incomplete preference less useful, and they gave clear feedback to the AI.

7.3 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.



(a) Text Summarization

(b) Photo Color Enhancement

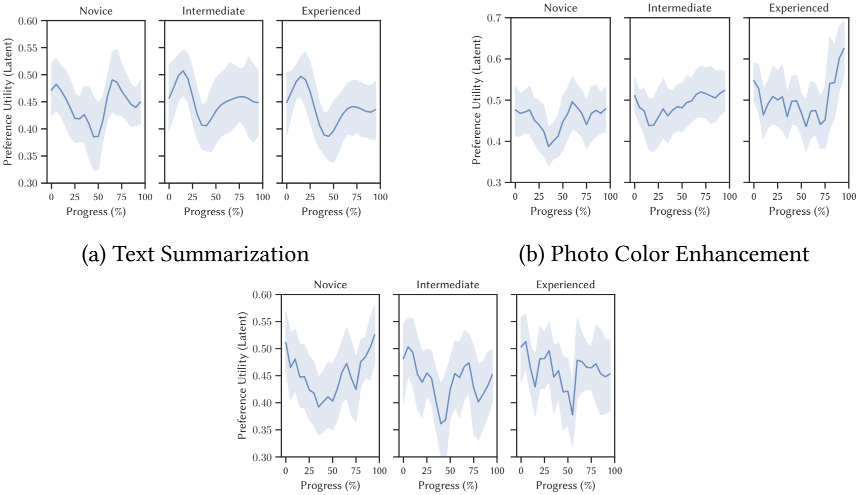
(c) 3D Model Simplification

Figure 7.5: Directly measured preference utility: The utility are normalized from rating labels (Terrible to Excellent). The results indicate regardless of the involved expertise, participants behave consistently, and in later iterations, the final ranking utility is higher than at the beginning of human-in-the-loop optimization.

7.3.1 Measured and Learned Preference Ranking Utility

As shown in Figure 7.5, for directly measured preference utility from ranking data, we fitted a linear mixed model [5, 88] (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, 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 human-in-the-loop optimization.

Regarding the learned latent utility from the BO optimizer, as illustrated in Figure 7.6, 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, 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.



(a) Text Summarization

(b) Photo Color Enhancement

(c) 3D Model Simplification

Figure 7.6: Learned latent preference utility: The inferred utility from the machine side (i.e., Bayesian optimization). Our results indicate provided ranking data from experienced participants are more consistent and effective in the learning process for the BO optimizer to learn than ranking data from novices.

7.3.2 Objective Outcome Quality

We normalized the iteration sequence and visualized the exploration progress in Figure 7.7. For analyzing the progress, we fitted linear mixed models for all metrics 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, t(1192) = 0.13, p = .894) and the effect of iteration is statistically non-significant and negative (β = -0.32, 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.

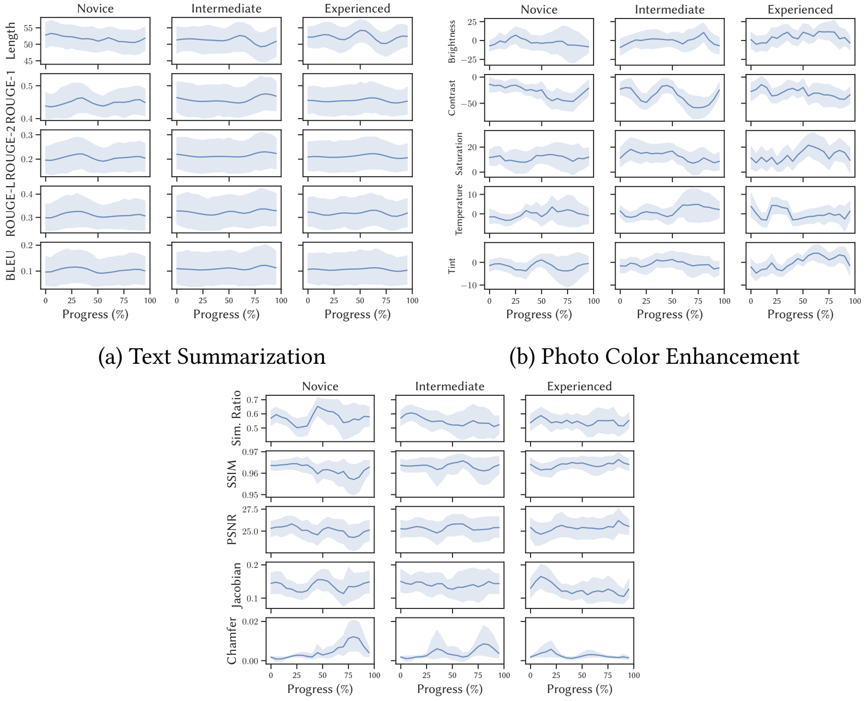
Lastly, for 3D model simplification, we found that experienced participants (β = 0.003, $CI_{95\%}$=[0.0001, 0.007], t(1192) = 2.00, p = .046) outperformed novices ($CI_{95\%}$=[-0.002, 0.002], t(1192) = 0.10) only in keeping surface distance low, meaning better in maintaining surface quality. We did 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 subsection 3.2.2. However, novices can achieve expert-level performance under the human-in-the-loop optimization context, similar to other contexts.

7.4 Discussion and Implications

The results in section 7.2 and section 7.3 can 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.

7.4.1 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* [118] 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*.



(a) Text Summarization

(b) Photo Color Enhancement

(c) 3D Model Simplification

Figure 7.7: System Outcomes' Objective Quality: Each context measured five metrics to the outcomes. Experts can identify technical differences compared to novices, such as minimizing Chamfer distance in 3D model simplification. Results indicate that novices produce expert-level performance in objective quality.

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 $[0,1]^r (s \in \mathbf{N}^+)$, and $\mathcal{O}$ be the outcome space generated from the parameter space. Then, the rational component of a human-in-the-loop optimization is to explore the outcome space $\mathcal{O}$ concerning a given set of objective metrics $\mathcal{M}_t(t \in \mathbf{N}^+)$. 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 human-in-the-loop 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 [28, 134], 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.

7.4.2 Expertise and Satisficing Decision Strategy

Based on the analysis of the outcome quality from the human-in-the-loop 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 human-in-the-loop 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 human-in-the-loop optimization.

7.4.3 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 human-in-the-loop 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 human-in-the-loop 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 [22][3], states that with an infinite number of events, the probability[4] 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 human-in-the-loop system can produce desired outcomes. In other words, this theoretical fact endorses that a sufficient amount of expertise

[3] In proposition 10.2.2 (b).

[4] Strictly speaking, the event happens *almost surely* as the Lebesgue measure is 1.

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.

7.4.4 Limitations

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 the 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 human-in-the-loop 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.

7.5 Summary

To zoom out from the detailed study and implications of the results, this chapter revealed empirical observations on the impact of involved expertise on the human-in-the-loop optimization systems, compared by three domain-specific examples. The results presented in this chapter inform us that the overall outcome quality is eventually not influenced by the involved user expertise in an optimization context. As an interpretation, this thesis argues that the human-in-the-loop system's outcomes are bound to the underlying machine algorithm's Pareto front. Based on the concept of Pareto efficiency in the definition of rationality and satisficing decision strategy in the definition of bounded rationality, one can consider the user objective alignment process moves along the front depending on the core values of the user. Combining with the observations in chapter 6, increasing expertise contributed to the behavior of exploring the Pareto front of the machine's intelligence. However, exploring the front does not objectively improve the machine outcomes and may even downgrade a user's overall satisfaction due to cognitive effects.

The observations reported in this chapter may be surprising, but they also leave us huge room for reflection and open the door for future work. We will look into them in the next final chapter.

8
Reflections and Outlook

The absurd is the essential concept and the first truth.

– Albert Camus, *The Myth of Sisyphus,* 1942

In the end, we reflect, starting with these questions: How reliable are our results? Why do human-in-the-loop optimization systems remain exciting and worth investigating despite our observations? Can rational machine programs truly benefit from bounded rational human feedback and align to achieve their objectives? This chapter discusses the findings philosophically and then outlines concrete future work.

8.1 Reflection

Behaviorism argues that objective decisions and judgments of a human can be observed through their behavior. However, this observation can be limited by the ambiguity of interpretation [144] due to a lack of observation and impermeable states of mind [116].

Despite this limitation, in chapter 5, we overcame the challenge and explored the opinion measurement interfaces as a complement to behaviorism [109] that measures “tendencies to evaluate an entity with some degree of favor or disfavor, ordinarily expressed in cognitive, affective, and behavioral responses” [34], which reflect human preferences or decisions under uncertainty. We showed incomplete and indifferent preferences commonly exist in human-in-the-loop optimization. These two indicators demonstrate unrevealed preferences and are often overlooked when designing machine optimizers to align with users' objectives. Does it mean we can adjust the design of optimizers to align user preferences inside a human's mind better?

To answer this question. We can go through a thought experiment: We designed and engineered a perfect machine optimizer based on all possible observations of the user's preferential behavior, and it can progressively and eventually align with the user preferences in their mind precisely in a few finite steps. Then, imagine a user who aims to trick the system and always provides *random* choices regardless of the outcome to the machine. What kind of user preference will the machine align to? What kind of outcome will the system generate? What does this optimization really mean? Can we assume the observed preferential choices are subject to true randomness even if the human user aims to provide "random choice"? Where is the source of this true randomness? Are we certain that this randomness is not influenced by their prior experiences subconsciousnessly?

Subjectively, as individual human beings, on request, we might always be able to *explain our own behavior in words.* Psychological research interprets the choice as a result of subjective confirmation bias [107] where people tend to rationalize their behavior. In turn, in a human-in-the-loop optimization system, the two involved entities might consistently suffer from this chain of suspicion and information asymmetry because each of them is trying to interpret the other's response. From each other's perspective, the behavior always encodes the freedom of will and the uncertain nature of the mind; interpretation constantly meets ambiguity, and the alignment process will continue.

Following this thought, even if we considered enough objective evaluating criteria in the optimizer and support *users to express on these different dimensions in the UI*, the subjective judgment of an individual may prefer a different ranking priority than another individual regarding concerning objectives. In a given context, to express a decision, a decision maker can select a partial set of objectives out of many objectives without informing any other objects. From the observer's perspective, one may interpret this decision-maker as having made an irrational choice, as there is neither sufficient interpretation of a decision nor enough observed information for the observer to infer the reasoning process within an objective mind. This essentially provides a counter-argument against the *Laplace's demon* [84] – a hypothetical superintelligent being could theoretically predict and calculate the future and past of the entire universe – because a human-in-the-loop optimization systems involves more than one intelligence and the interaction between them is non-deterministic and cannot be fully inferred.

Is this a dead end if we believe that we are doomed to fail in designing rational machines with perfect algorithms to optimize system outcomes and align them with users? What else can we do? Beyond the provided design recommendations in chapter 6 that mitigate the problem, a more general approach to end such an objective alignment process efficiently is to meet the following: 1) a shared common goal and both parties need to adapt rather than unconditionally requiring the adaptation of one to the other, especially if the other might not have concrete and revealed preference. Otherwise, the optimization process does not have sufficient feedback to optimize; 2) transparently sharing the same and complete information. Otherwise, when zooming out from a detailed inspection of different concerning factors, the observation will also be limited by the central limited theorem, where their findings become uninterpretable or ambiguous because there are too many confounding variables; 3) the involved two entities use sufficient rationality to minimize to an optimal gap at every each step because optimization errors can occur due to suboptimal strategy.

In this light, we also need to rethink the relationship between rationality and our intelligence. Preference logic shows a definition by relying on assumptions of completeness and transitivity, which is a way to describe rationality. Similarly, the fundamental axioms of probability theory can also be considered as a definition to address rationality (cf. Dutch book argument [122], a violation of probability axioms license senseless and contradictory behavior). Therefore, conventionally, rational assumption implicitly suggests that it presents our intelligence. However, in the empirical observations of this thesis, we have seen users often violating these assumptions when interacting with a system, not only by pure mistakes but rather by being framed as a misalignment between machine objectives and user objectives or pure creativity. Should we do a better design to help humans be more rational and let machines understand us better? Is it our objective to eliminate the "irrational" component? What does it mean for us if we behave purely rationally and machines can quickly learn and adapt to it? Can we consider the machine accumulated our intelligence? What can we interpret if the individual user mismatches the machine's intelligence? What truly shaped their intelligence as well as ours?

Ultimately, physical time limits the freedom of our minds. We ask ourselves eventually: how can we better keep our intelligence in an interaction loop? The answer I found is to seek unexperienced experiences. From this process, the interaction of rationality that is learned from others and creativity that originates from ourselves together shapes our intelligence and empowers us to further create essence to existence.

8.2 Future Work

Looking ahead, the above reflection contributes the following concrete ideas for continuing the work on the research of human-in-the-loop systems.

Involving unstructured and unaggregated human feedback: Our considered design space of opinion measurement interfaces in chapter 5 only covers structured human inputs bound to suitable optimizers, i.e., utility and preferential data. With the increasing developments in large language models, it would also be interesting and challenging to explore, quantify, and evaluate human-in-the-loop interfaces that permit unstructured inputs such as using a text prompt[1] to alter media data using diffusion models [129]. The results presented in chapter 5 provided a starting point in this direction. In addition, all the empirical explorations in this thesis ask humans to provide aggregated feedback, which may need to be more clear from the interaction process. Another possible future exploration is to dig more into a more granular level of feedback from different judging criteria.

Modeling indifference and incomplete preference: In our studies, we designed the user interface to distinguish between two possible interpretations (ignorance or uncurious) of incomplete preference. Still, the optimizers discarded the collected indifference and incomplete preferences because the underlying optimizer did not support them. However, indifference and incomplete preference are common in human decision-making [27]. In chapter 5 and chapter 7, we showed that although users rarely express indifference and incomplete preferences, understanding them may help us better in the optimization process towards user expectation. Although we have seen initial work (e.g., [105, 108]) from both the machine learning community and economic research, A closer look into understanding and modeling indifference and incomplete preferences will be fundamentally challenging and interesting to go.

Simulating human priors and feedback: When conducting a research experiment, creating a feedback simulator, either random or with a particular assumed human prior [59], to interact with a human-in-the-loop system may be helpful for researchers to understand whether the outcome based on human feedback is better than a reproducible baseline. Yet, the simulation approach is not widely accepted in the HCI community due to the complexity of humans. However, we have seen promising trend advocates [103] for applying them, and we believe it will be a promising and valid approach for future HCI research.

[1] `https://openai.com/blog/chatgpt/`, *last accessed 17.02.2023*

Understanding human behaviors on the Pareto front: One of the essential ideas regarding human-in-the-loop systems is to adapt user expectations progressively. Such an adaptation happens based on the measure of performance space and objective functions designed within an optimizer. Suppose a machine can measure the objective and infer the next optimal based on simulated results with an optimization process. In that case, the outcomes towards more to the Pareto front are determined by a set of fixed objectives when designing the machine. Does it make more sense for the machine systems to jump to the front and produce optimal outcomes without human inputs? Is the whole adaptation idea still sensible to involve humans in the optimization phase? What would change if an involved human constantly faces Pareto optimal results and the feedback loop only adjusts to the system outcomes on the front? Although one of the direct technical challenges is to find the Pareto front, these questions are worth exploring.

Exploring mismatches between individual and collective intelligence: When individual intelligence cannot compete with essentially embedded collective intelligence in the loop, how can we discover the gaps and mismatches between individual intelligence and the machine's intelligence? How can we better support individual users to explore the space created by collective intelligence? How can we design a system to help users understand how well they have explored the space fixed by the current machine system? This is also yet another interesting direction to investigate.

Bibliography

[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] Robert L. Armstrong. 1987. The Midpoint on a Five-Point Likert-Type Scale. *Perceptual and Motor Skills.* 64, 2 (Apr 1987), 359–362. https://doi.org/10.2466/pms.1987.64.2.359

[3] Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G 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' 2020)*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.). Curran Associates, Inc., Red Hook, NY, USA, Article 1807, 15 pages. https://dl.acm.org/doi/10.5555/3495724.3497531

[4] 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.* Technical Report. SRI International Menlo Park California Artificial Intelligence Center. https://apps.dtic.mil/sti/citations/ADA458355

[5] 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 arXiv:1406.5823 [stat.CO]

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

[7] 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 (Sep 2013), 51–76. `https://doi.org/doi.org/10.1111/cgf.12014`

[8] Mario Botsch, Leif Kobbelt, Mark Pauly, Pierre Alliez, and Bruno Levy. 2010. *Polygon Mesh Processing.* CRC Press, New York, NY, USA. pp. 111–130. `https://doi.org/10.1201/b10688`

[9] Lyle E. Bourne Jr., James A. Kole, and Alice F. Healy. 2014. Expertise: defined, described, explained. *Frontiers in Psychology.* 5, Article 186 (Mar 2014), 3 pages. `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* (Vancouver, British Columbia, Canada) *(NIPS'07).* Curran Associates Inc., Red Hook, NY, USA, 409–416. `https://dl.acm.org/doi/abs/10.5555/2981562.2981614`

[13] Eric Brochu, Abhijeet Ghosh, and Nando de Freitas. 2007. Preference Galleries for Material Design. In *ACM SIGGRAPH 2007 Posters* (San Diego, California) *(SIGGRAPH '07).* Association for Computing Machinery, New York, NY, USA, 2 pages. `https://doi.org/10.1145/1280720.1280834`

[14] Daniel Buschek, Lukas Mecke, Florian Lehmann, and Hai Dang. 2021. Nine Potential Pitfalls when Designing Human-AI Co-Creative Systems. `https://doi.org/10.48550/arXiv.2104.00358` arXiv:2104.00358 [cs.HC]

[15] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO, USA) *(CVPR 2011)*. IEEE, New York, NY, USA, 97–104. `https://doi.org/10.1109/CVPR.2011.5995413`

[16] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. `https://doi.org/10.1145/3290605.3300234`

[17] Juan C. Caicedo, Ashish Kapoor, and Sing Bing Kang. 2011. Collaborative personalization of image enhancement. In *Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO, USA). IEEE, New York, NY, USA, 249–256. `https://doi.org/10.1109/CVPR.2011.5995439`

[18] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to Rank: From Pairwise Approach to Listwise Approach. In *Proceedings of the 24th International Conference on Machine Learning* (Corvalis, Oregon, USA) *(ICML '07)*. Association for Computing Machinery, New York, NY, USA, 129–136. `https://doi.org/10.1145/1273496.1273513`

[19] Toby Chong, I-Chao Shen, Issei Sato, and Takeo Igarashi. 2021. Interactive Optimization of Generative Image Modelling using Sequential Subspace Search and Content-based Guidance. *Computer Graphics Forum.* 40, 1 (Feb 2021), 279–292. `https://doi.org/10.1111/cgf.14188`

[20] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4302–4310. `https://dl.acm.org/doi/abs/10.5555/3294996.3295184`

[21] Wei Chu and Zoubin Ghahramani. 2005. Preference Learning with Gaussian Processes. In *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany) *(ICML '05)*. Association for Computing Machinery, New York, NY, USA, 137–144. `https://doi.org/10.1145/1102351.1102369`

[22] Donald L. Cohn. 2013. *Probability*, In *Measure Theory: Second Edition*. Springer, New York, NY, USA. pp. 307–371. https://doi.org/10.1007/978-1-4614-6956-8_10

[23] 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* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 293–297. https://doi.org/10.1145/3340631.3394883

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

[25] M. Corsini, M. C. Larabi, K. Wang, and et al. 2013. Perceptual Metrics for Static and Dynamic Triangle Meshes. *Computer Graphics Forum*. 32, 1 (Jan 2013), 101–125. https://doi.org/10.1111/cgf.12001

[26] Harold P. Van Cott and Robert G. Kinkade. 1972. *Man-Machine Dynamics*, In *Human Engineering Guide to Equipment Design*. American Institutes for Research, Washington, D.C. pp. 229–230.

[27] Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. 2020. Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective 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 826, 14 pages. https://dl.acm.org/doi/10.5555/3495724.3496550

[28] Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. 2021. Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement. In *Proceedings of the 35th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada), 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

[29] 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

[30] Ye Ding, Myunghee Kim, Scott Kuindersma, and Conor J. Walsh. 2018. Human-in-the-loop optimization of hip assistance with a soft exosuit during walking. *Science Robotics*. 3, 15 (Feb 2018), 8 pages. https://doi.org/10.1126/scirobotics.aar5438

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

[32] John J. Dudley, Jason T. Jacques, and Per Ola Kristensson. 2019. Crowdsourcing Interface Feature Design with Bayesian Optimization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300482

[33] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems*. 8, 2, Article 8 (Jun 2018), 37 pages. https://doi.org/10.1145/3185517

[34] Alice H. Eagly and Shelly Chaiken. 1993. *The psychology of attitudes.* Harcourt Brace Jovanovich College Publishers, Orlando, FL, US. pp. XXII, 794.

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

[36] 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* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 96–106. https://doi.org/10.1145/3301275.3302262

[37] Joyce J. Elam and Melissa Mead. 1990. Can Software Influence Creativity? *Information Systems Research*. 1, 1 (Mar 1990), 22 pages. https://doi.org/10.1287/isre.1.1.1

[38] Ziv Epstein, Aaron Hertzmann, the Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R. Frank, Matthew

Groh, Laura Herman, Neil Leach, Robert Mahari, Alex "Sandy" Pentland, Olga Russakovsky, Hope Schroeder, and Amy Smith. 2023. Art and the science of generative AI. *Science.* 380, 6650 (2023), 1110–1111. https://doi.org/10.1126/science.adh4451 arXiv:https://www.science.org/doi/pdf/10.1126/science.adh4451

[39] 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* (Utrecht, Netherlands) *(UMAP '21).* Association for Computing Machinery, New York, NY, USA, 29–34. https://doi.org/10.1145/3450614.3461683

[40] Johannes Fürnkranz and Eyke Hüllermeier. 2003. Pairwise Preference Learning and Ranking. In *Proceedings of the 14th European Conference on Machine Learning* (Cavtat-Dubrovnik, Croatia) *(ECML'03).* Springer-Verlag, Berlin, Heidelberg, 145–156. https://doi.org/10.1007/978-3-540-39857-8_15

[41] 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 (SIGGRAPH '97).* ACM Press/Addison-Wesley Publishing Co., USA, 209–216. https://doi.org/10.1145/258734.258849

[42] 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* (Research Triangle Park, North Carolina, USA) *(VIS '98).* IEEE Computer Society Press, Washington, DC, USA, 263–269. https://dl.acm.org/doi/10.5555/288216.288280

[43] 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 (Mar 2009), 93–105. https://doi.org/10.1080/14639220802059190

[44] 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

[45] 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, 1282–1291. https://doi.org/10.5555/3305381.3305514

[46] Miriam Greis, Hendrik Schuff, Marius Kleiner, Niels Henze, and Albrecht Schmidt. 2017. Input Controls for Entering Uncertain Data: Probability Distribution Sliders. *Proceedings of the ACM on Human-Computer Interaction.* 1, EICS, Article 3 (Jun 2017), 17 pages. https://doi.org/10.1145/3095805

[47] 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, Article 167 (Jun 1996), 293–323 pages. https://doi.org/10.1037/1076-8971.2.2.293

[48] 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

[49] Robert T. Gschwind and Irving L. Chidsey. 1981. *Analysis of Man-in-the-Loop Control Systems in the Presence of Nonlinearities.* Technical Report. Army Ballistic Research Lab Aberdeen Providing Ground Md. https://apps.dtic.mil/sti/citations/ADA102574

[50] Daniel M. Hausman. 2011. *Preference, Value, Choice, and Welfare.* Cambridge University Press, New York, USA. pp. 1–10.

[51] 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

[52] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. https://doi.org/10.48550/arXiv.1904.09751 arXiv:1904.09751 [cs.CL]

[53] 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 (Lecture Notes in Computer Science)*. Springer, Cham, 81–95. https://doi.org/10.1007/978-3-319-45507-5_6

[54] Hugues Hoppe. 1996. Progressive Meshes. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '96)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/237170.237216

[55] 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

[56] 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

[57] 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 (Feb 2006), 143–150. https://doi.org/10.1111/j.1467-9280.2006.01677.x

[58] Wenzel Jakob, Marco Tarini, Daniele Panozzo, and Olga Sorkine-Hornung. 2015. Instant Field-Aligned Meshes. *ACM Transactions on Graphics.* 34, 6, Article 189 (Nov 2015), 15 pages. https://doi.org/10.1145/2816795.2818078

[59] Edwin T. Jaynes. 1968. Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics.* 4, 3 (Sep 1968), 227–241. https://doi.org/10.1109/TSSC.1968.300117

[60] Daniel Kahneman. 2011. *Thinking, Fast and Slow.* Penguin Random House UK. pp. 19–97.

[61] 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 (Mar 1991), 193–206. https://doi.org/10.1257/jep.5.1.193

[62] Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise.* William Collins, UK. pp. 11–94.

[63] 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

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

[65] Saikishore Kalloori, Francesco Ricci, and Rosella Gennari. 2018. Eliciting Pairwise Preferences in Recommender Systems. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18).* Association for Computing Machinery, New York, NY, USA, 329–337. https://doi.org/10.1145/3240323.3240364

[66] 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. https://dl.acm.org/doi/10.5555/2343576.2343643

[67] 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* (Denver, Colorado, USA) *(CHI '17).* Association for Computing Machinery, New York, NY, USA, 1295–1306. https://doi.org/10.1145/3025453.3025576

[68] 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

[69] Jakob Karolus and Albrecht Schmidt. 2018. Proficiency-Aware Systems: Adapting to the User's Skills and Expertise. In *Proceedings of the 7th ACM International Symposium on Pervasive Displays* (Munich, Germany) *(PerDis '18).* Association for Computing Machinery, New York, NY, USA, Article 33, 2 pages. https://doi.org/10.1145/3205873.3210708

[70] Jakob Karolus and Paweł W. Woźniak. 2021. Proficiency-aware systems: Designing for user reflection in context-aware systems. *Information Technology.* 63, 3 (Jul 2021), 167–175. `https://doi.org/10.1515/itit-2020-0039`

[71] George Karypis and Vipin Kumar. 1997. *METIS: A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices.* Technical Report. University of Minnesota, Department of Computer Science / Army HPC Research Center. `http://conservancy.umn.edu/handle/11299/215346`

[72] Myunghee Kim, Ye Ding, Philippe Malcolm, Jozefien Speeckaert, Christoper J. Siviy, Conor J. Walsh, and Scott Kuindersma. 2017. Human-in-the-loop Bayesian optimization of wearable device parameters. *PLOS ONE.* 12, 9 (Sep 2017), 15 pages. `https://doi.org/10.1371/journal.pone.0184054`

[73] 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* (San Jose, California, USA) *(CHI '16).* Association for Computing Machinery, New York, NY, USA, 2390–2395. `https://doi.org/10.1145/2858036.2858402`

[74] Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. 2021. Explainable Artificial Intelligence for Human Decision-Support System in Medical Domain. `https://doi.org/10.48550/arXiv.2105.02357` arXiv:2105.02357 [cs.HC]

[75] 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 (Apr 2000), 401–420. `https://doi.org/10.1002/(SICI)1097-0207(20000530)48:3<401::AID-NME880>3.0.CO;2-D`

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

[77] 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. In *Proceedings of the 2019 CHI Conference on Hu-*

man Factors in Computing Systems (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[78] 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* (Melbourne, VIC, Australia). IEEE, New York, NY, USA, 37–38. https://doi.org/10.1109/QoMEX.2012.6263880

[79] Ben Kotzee and Jp Smit. 2018. *Two Social Dimensions of Expertise*, In *Education and Expertise*. John Wiley & Sons, Ltd, New Jersey, USA. 99–116 pages. https://doi.org/10.1002/9781119527268.ch5

[80] 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* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 65–74. https://doi.org/10.1145/2642918.2647386

[81] 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

[82] Yuki Koyama, Issei Sato, and Masataka Goto. 2020. Sequential gallery for interactive visual design optimization. *ACM Transactions on Graphics.* 39, 4, Article 88 (Jul 2020), 12 pages. https://doi.org/10.1145/3386569.3392444

[83] Yuki Koyama, Issei Sato, Daisuke Sakamoto, and Takeo Igarashi. 2017. Sequential line search for efficient visual design optimization by crowds. *ACM Transactions on Graphics.* 36, 4, Article 48 (Jul 2017), 11 pages. https://doi.org/10.1145/3072959.3073598

[84] Boris Kožnjak. 2015. Who let the demon out? Laplace and Boscovich on determinism. *Studies in History and Philosophy of Science Part A.* 51 (2015), 42–52. https://doi.org/10.1016/j.shpsa.2015.03.002

[85] Markus Krause and Jan Smeddinck. 2011. Human computation games: A survey. In *2011 19th European Signal Processing Conference* (Vancouver, BC,

Canada). IEEE, New York, NY, USA, 754–758. https://ieeexplore.ieee.org/abstract/document/7074262

[86] 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

[87] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China). Association for Computational Linguistics, 540–551. https://doi.org/10.18653/v1/D19-1051

[88] 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 (Dec 2017), 26 pages. https://doi.org/10.18637/jss.v082.i13

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

[90] 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 (Jan 2012), 151–163. https://doi.org/10.1111/j.1756-8765.2011.01175.x

[91] 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

[92] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out* (Barcelona, Spain). Association for Computational Linguistics, 74–81. https://aclanthology.org/W04-1013

[93] 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 Conference, 4235–4258. https://proceedings.mlr.press/v151/jerry-lin22a.html

[94] Steson Lo and Sally Andrews. 2015. To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology.* 6, Article 1171 (Aug 2015), 16 pages. https://doi.org/10.3389/fpsyg.2015.01171

[95] Wai-Lan Luk. 1994. *Multi-user interface for group ranking: a user-centered approach.* Ph. D. Dissertation. University of British Columbia. https://doi.org/10.14288/1.0087518

[96] 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

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

[98] Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. 2016. The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales. 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, 5421–5432. https://doi.org/10.1145/2858036.2858063

[99] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (Barcelona, Spain), Dekang Lin and Dekai Wu (Eds.). Association for Computational Linguistics, 404–411. https://aclanthology.org/W04-3252

[100] Petrus Mikkola, Milica Todorović, Jari Järvi, Patrick Rinke, and Samuel Kaski. 2020. Projective Preferential Bayesian Optimization. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. PMLR, MLResearchPress, Article 639, 9 pages. `https://dl.acm.org/doi/abs/10.5555/3524938.3525577`

[101] Robert M. Monarch. 2021. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI.* Simon and Schuster, USA. pp. 3–22.

[102] J. Močkus. 1975. *On Bayesian Methods for Seeking the Extremum*, In *Optimization Techniques IFIP Technical Conference.* Springer, Berlin, Heidelberg. pp. 400–404. `https://doi.org/10.1007/978-3-662-38527-2_55`

[103] Roderick Murray-Smith, Antti Oulasvirta, Andrew Howes, Jörg Müller, Aleksi Ikkala, Miroslav Bachinski, Arthur Fleig, Florian Fischer, and Markus Klar. 2022. What Simulation Can Do for HCI Research. *Interactions.* 29, 6 (Nov 2022), 48–53. `https://doi.org/10.1145/3564038`

[104] Wolfgang Neubarth. 2010. *Drag & drop: A flexible method for moving objects, implementing rankings, and a wide range of other applications*, In *Advanced Methods for Conducting Online Behavioral Research.* American Psychological Association, Washington, DC, US. pp. 63–74. `https://doi.org/10.1037/12076-005`

[105] Quoc Phong Nguyen, Sebastian Tay, Bryan Kian Hsiang Low, and Patrick Jaillet. 2021. Top-k Ranking Bayesian Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence.* 35, 10 (May 2021), 9135–9143. `https://doi.org/10.1609/aaai.v35i10.17103`

[106] Tien T. Nguyen, Daniel Kluver, Ting-Yu Wang, Pik-Mai Hui, Michael D. Ekstrand, Martijn C. Willemsen, and John Riedl. 2013. Rating Support Interfaces to Improve User Experience and Recommender Accuracy. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13).* Association for Computing Machinery, New York, NY, USA, 149–156. `https://doi.org/10.1145/2507157.2507188`

[107] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology.* 2, 2 (Jun 1998), 175–220. `https://doi.org/10.1037/1089-2680.2.2.175`

[108] Kirby Nielsen and Luca Rigotti. 2022. Revealed Incomplete Preferences. https://doi.org/10.48550/arXiv.2205.08584 arXiv:2205.08584v3 [econ.GN]

[109] Syavash Nobarany, Louise Oram, Vasanth Kumar Rajendran, Chi-Hsiang Chen, Joanna McGrenere, and Tamara Munzner. 2012. The Design Space of Opinion Measurement Interfaces: Exploring Recall Support for Rating and Ranking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2035–2044. https://doi.org/10.1145/2207676.2208351

[110] Marc Olano, Bob Kuehne, and Maryann Simmons. 2003. Automatic Shader Level of Detail. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware* (San Diego, California) *(HWWS '03)*. Eurographics Association, Goslar, DEU, 7–14. https://dl.acm.org/doi/10.5555/844174.844176

[111] 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:2109.08183 [cs.HC]

[112] 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* (Darmstadt, Germany) *(MuC '22)*. Association for Computing Machinery, New York, NY, USA, 158–168. https://doi.org/10.1145/3543758.3543761

[113] Changkun Ou, Sven Mayer, Daniel Buschek, and Andreas Butz. 2024. Rethinking Opinion Measurement Interfaces for Human-in-the-loop Optimization. *ACM ACM Transactions on Computer-Human Interaction.* (2024), 30 pages. SUBMITTED

[114] Changkun Ou, Sven Mayer, and Andreas Martin Butz. 2023. The Impact of Expertise in the Loop for Exploring Machine Rationality. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 307–321. https://doi.org/10.1145/3581641.3584040

[115] 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 (Chengdu, Sichuan, China) *(ICAIBD' 19)*. IEEE, 369–378. https://doi.org/10.1109/ICAIBD.2019.8836972

[116] Søren Overgaard. 2006. The Problem of Other Minds: Wittgenstein's Phenomenological Perspective. *Phenomenology and the Cognitive Sciences.* 5, 1 (Mar 2006), 53–73. https://doi.org/10.1007/s11097-005-9014-7

[117] 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 on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[118] Vilfredo Pareto. 1912. *Manuel d'économie politique.* Vol. 18. pp. 3.

[119] John W Payne, John William Payne, James R Bettman, and Eric J Johnson. 1993. *The adaptive decision maker.* Cambridge University Press. pp. 248–263.

[120] 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 (Jul 2021), 17 pages. https://doi.org/10.1145/3450626.3459941

[121] Federico Ponchio. 2008. *Multiresolution structures for interactive visualization of very large 3D datasets.* Ph. D. Dissertation. Clausthal University of Technology. http://d-nb.info/997062789/34

[122] Joel Pust. 2021. Dutch Books and Logical Form. *Philosophy of Science.* 88, 5 (Dec 2021), 961–970. https://doi.org/10.1086/714997

[123] 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

[124] 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

[125] Carl Edward Rasmussen, Christopher K. I. Williams, and Francis Bach. 2004. *Gaussian Processes in Machine Learning*, In *Advanced Lectures on Machine Learning: ML Summer Schools 2003.* Springer, Berlin, Heidelberg. pp. 63–71. https://doi.org/10.1007/978-3-540-28650-9_4

[126] Nornadiah Mohd Razali, Yap Bee Wah, et al. 2011. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics.* 2, 1 (Jan 2011), 21–33.

[127] 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

[128] 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 Machine Learning*. PMLR, Beijing, China, 1278–1286. https://proceedings.mlr.press/v32/rezende14.html

[129] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, New Orleans, USA, 10684–10695. https://doi.org/10.48550/arXiv.2112.10752

[130] 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

[131] Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review.* 16, 2 (Apr 2009), 225–237. https://doi.org/10.3758/PBR.16.2.225

[132] David Schmidtz. 2004. *Satisficing as a Humanly Rational Strategy*, In *Satisficing and Maximizing: Moral Theorists on Practical Reason.* Cambridge University Press, New York, USA. 30–59 pages. https://philpapers.org/rec/SCHSAA-18

[133] Matthias Schonlau, William J. Welch, and Donald R. Jones. 1998. *Global versus Local Search in Constrained Optimization of Computer Models*, In *New Developments and Applications in Experimental Design (Lecture Notes - Monograph Series).* Vol. 34. Institute of Mathematical Statistics, Hayward, California. 11–25 pages. http://www.jstor.org/stable/4356058

[134] Adriana Schulz, Harrison Wang, Eitan Grinspun, Justin Solomon, and Wojciech Matusik. 2018. Interactive Exploration of Design Trade-Offs. *ACM Transactions on Graphics.* 37, 4, Article 131 (Jul 2018), 14 pages. https://doi.org/10.1145/3197517.3201385

[135] 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

[136] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. https://doi.org/10.48550/arXiv.1704.04368 arXiv:1704.04368 [cs.CL]

[137] 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

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

[139] Mark Sherer, Paula Bergloff, Corwin Boake, Walter High Jr, and Ellen Levin. 1998. The Awareness Questionnaire: factor structure and internal consistency. *Brain Injury.* 12, 1 (Jan 1998), 63–68. https://doi.org/10.1080/026990598122863

[140] 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

[141] 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

[142] Herbert A. Simon. 1977. *Scientific Discovery and the Psychology of Problem Solving*, In *Models of Discovery: And Other Topics in the Methods of Science*. Springer Netherlands, Dordrecht, 286–303. https://doi.org/10.1007/978-94-010-9521-1_16

[143] 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 (Dec 2020), 759–775. https://doi.org/10.1162/tacl_a_00344

[144] B. F. Skinner. 2011. *About Behaviorism*. Knopf Doubleday Publishing Group, New York, NY, USA. pp. 106–194.

[145] Ralph H. Sprague. 1980. A Framework for the Development of Decision Support Systems. *MIS Quarterly*. 4, 4 (Dec 1980), 26 pages. https://doi.org/10.2307/248957

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

[147] Jeffrey W. Treem and Paul M. Leonardi. 2016. *Expertise, Communication, and Organizing*. Oxford University Press, New York, NY, USA. pp. 1–5.

[148] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science*. 185, 4157 (Sep 1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

[149] 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

[150] 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*

(Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376877

[151] Luis von Ahn. 2007. Human Computation. In *Proceedings of the 4th International Conference on Knowledge Capture* (Whistler, BC, Canada) *(K-CAP '07)*. Association for Computing Machinery, New York, NY, USA, 5–6. https://doi.org/10.1145/1298406.1298408

[152] Eric-Jan Wagenmakers. 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*. 14, 5 (Oct 2007), 779–804. https://doi.org/10.3758/BF03194105

[153] 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

[154] 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)*. The Eurographics Association, Goslar, DEU, 33–37. https://doi.org/10.2312/SGP/SGP07/033-037

[155] 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* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 243–253. https://doi.org/10.1145/3377325.3377509

[156] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural Text Generation With Unlikelihood Training. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview, Addis Ababa, Ethiopia, 18 pages. https://doi.org/10.48550/arXiv.1908.04319

[157] Michael Wilber, Iljung Kwak, and Serge Belongie. 2014. Cost-Effective HITs for Relative Similarity Comparisons. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. 2, 1 (Sep 2014), 227–233. https://doi.org/10.1609/hcomp.v2i1.13152

[158] 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 arXiv:1712.00424 [stat.ML]

[159] Dennis Wixon and John Whiteside. 1985. Engineering for Usability (Panel Session): Lessons from the User Derived Interface. *ACM SIGCHI Bulletin.* 16, 4 (Apr 1985), 144–147. https://doi.org/10.1145/1165385.317484

[160] 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

[161] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. *Interactions.* 23, 3 (Jun 2016), 38–44. https://doi.org/10.1145/2907069

[162] 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* (Honolulu, HI, USA) *(CHI EA '20).* Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382821

[163] 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* (College Station, TX, USA) *(IUI '21).* Association for Computing Machinery, New York, NY, USA, 43–47. https://doi.org/10.1145/3397481.3450663

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig und ohne unerlaubte Beihilfe angefertigt wurde.

München, den 17.02.2023

CHANGKUN OU

Changkun Ou

The Intelligence in the Loop: Empirical Explorations and Reflections

For decades, engineering in computing systems has used a human-in-the-loop servo mechanism. A conscious human being is usually believed, in a rational manner, to operate, assist, and control the machine to achieve desired objectives. Over time, researchers have started to use human-in-the-loop schemes in more abstract tasks, such as iterative interface design problems. However, with the observations and developments in social science, the underlying rationality assumption is strongly challenged, and humans make mistakes. With the recent advances in computer science regarding artificial intelligence, data-driven algorithms could achieve human-level performance in certain aspects. The human-in-the-loop mechanism is being reconsidered and reshaped towards an extended vision to assist human decision-making or creativity in the human-computer interaction (HCI) research field.

This dissertation explores the boundary for human-in-the-loop optimization systems to succeed and be beneficial, focusing on understanding an iterative interaction loop where machine agents are designed to interact with human beings that may behave using bounded rational policies, align objectives iteratively, and optimize the machine outcomes. We analyzed the building blocks in a human-in-the-loop optimization system and then designed three studies to assess each element, including user interfaces, preference optimizers, human expertise, cognitive effects, satisfaction, termination condition, etc. The presented observations and results eventually approached more fundamental questions regarding the definition of intelligence and suggested an answer on how we could succeed in keeping our *intelligence in the loop*.