The Intelligence in the Loop
Empirical Explorations and Reflections

Changkun Ou

Ph.D. Defense | Munich, Germany | April 3rd, 2023
Motivation
Low Rendering Cost

High Rendering Cost
Solution Space
Solution Space
Human As the System Operator

Solution Space
Human-in-the-Loop (HITL) Systems

- A human-in-the-loop system refers to a type of system or process that involves human intervention or oversight at some point during its operation [Wixon and Whiteside, CHI’ 85]
- The role of a human is to steer the system behavior, and handle intuition, creativity, and decision-making which the machine could not handle procedurally.
Human-in-the-Loop Optimization Systems

A human-in-the-loop optimization system refers to systems or processes that involve an underlying optimization process towards user expectation or preference.
HITL Optimization System Building Blocks
HITL Optimization System Building Blocks
HITL Optimization System Building Blocks

Objective

World

problem

context

user
HITL Optimization System Building Blocks

HITL Optimization System

Objective World

User

Inputs

Behaviors

Problem Context
HITL Optimization System Building Blocks

HITL Optimization System

Human Intelligence

Objective World

Problem context

inputs

behaviors

perceives

user
HITL Optimization System Building Blocks

Human Intelligence

- user
  - perceives
  - behaviors
  - inputs

Objective World

- problem context

renders
optimizes
characterizes
measures
performance
solution
space
user interface

World
Human Intelligence
Intelligence
HITL Optimization System Building Blocks

- **Human Intelligence**
  - Empirical Knowledge
  - User
    - User Interface Space
      - Inputs
        - Behaviors
        - Perceives

- **Machine Intelligence**
  - Performance Space
    - Measures
  - Solution Space
    - Optimizes
    - Characterizes
    - Renders

- **Objective World**

Problem Context
**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 these criteria?

**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?
Empirical Study I
The Human in the Infinite Loop

**Objective**

The infinite loop is a concept where actions and their consequences create a feedback loop, often leading to unexpected outcomes. This loop can be observed in various contexts, from software development to philosophical thought experiments.

**World**

The infinite loop can be understood in the context of human-computer interaction, where users interact with systems that exhibit loops. This interaction can lead to unpredictable behaviors and outcomes.

**Human Intelligence**

Understanding the infinite loop requires insights from cognitive psychology and computer science. The loop can be analyzed from different perspectives, such as user experience and system design.

**Behavior**

Behavioral models of the infinite loop can be complex and involve multiple factors. These models are often used to predict and manage the outcomes of actions within looped systems.

**Perceives**

Perceives are the human users who interact with the infinite loop. They may be unaware of the loop's existence or may intentionally create it as a tool for exploration or problem-solving.

**Inputs**

Inputs to the infinite loop can come from various sources, including user actions, system feedback, and external factors. These inputs can be processed by the system, leading to new outputs that feed back into the loop.

**Problem Context**

Understanding the infinite loop in the context of human-computer interaction is crucial for designing systems that are user-friendly and efficient. This understanding can help in creating better user experiences and more robust systems.

---

**Figure 1**: A human in the loop UI modeling system. A server generates differently processed variations of a complex UI model and disperses them to user interfaces, which process their variations as a UI states, which are then stored in a database. The figure shows a user interface with a server, a database, and an AI system, illustrating the interaction loop.
Gallery-based Approach

Examples

Text Editing [Terry et al. 2002] [Lee et al. 2010] [Simpson et al. 2020]
Photo Color Enhancement [Koyama et al. 2014, 2016]
Melody composition [Zhou et al. 2021]
Interface design [Liao et al. 2021a, 2021b]
Material design [Brochu et al. 2007]
Animation [Brochu et al. 2010]
Illumination [Marks et al. 1997]
Prototype & User Workflow

Ask for Feedback

Rates

Dispatches
[Hoppe et al. 1996]
[Garland and Heckbert 1998]
[Jakob et al. 2015]
Prototype & User Workflow

Ask for Feedback

Rates

Dispatches
[Hoppe et al. 1996]
[Garland and Heckbert 1998]
[Jakob et al. 2015]

Mesh Space

[Rates]
Prototype & User Workflow

Ask for Feedback

Rates

Dispatches

[Hoppe et al. 1996]
[Garland and Heckbert 1998]
[Jakob et al. 2015]

Mesh Space
Prototype & User Workflow

Ask for Feedback

Rates

Dispatches

[Hoppe et al. 1996]
[Garland and Heckbert 1998]
[Jakob et al. 2015]

Mesh Space
Prototype & User Workflow

Ask for Feedback

Rates

Dispatches

[Hoppe et al. 1996]
[Garland and Heckbert 1998]
[Jakob et al. 2015]
Prototype & User Workflow

Ask for Feedback

Rates

Dispatches

Bayesian Optimization

[Hoppe et al. 1996]
[Garland and Heckbert 1998]
[Jakob et al. 2015]

[Gonzalez et al. 2017]
Prototype & User Workflow

Ask for Feedback

Rates

Dispatches

Bayesian Optimization

Inference

Mesh Space

[Hoppe et al. 1996]
[Garland and Heckbert 1998]
[Jakob et al. 2015]
[Gonzalez et al. 2017]
Bayesian Optimization (BO)

Bayesian optimization [Mockus, 1978] aims to find \( x^* = \arg\max_{x \in \mathcal{X}} f(x) \)

With an initial dataset \( D_0 = \{(x_i, y_i)\}_{i=1}^{N} \)

1. Choose some prior measure over the space of possible objectives \( f(x) \sim GP(\mu(x), k(x, x')) \)

2. Combine prior and likelihood to get a posterior measure over objective given some observations

3. Use the posterior to infer next evaluation according to some acquisition function \( x_{t+1} = \arg\max_{x \in \mathcal{X}} AF(x; D_t) \)

4. Augment the data.

Iterate between 2 and 4 until the evaluation budget is over
Preferential Bayesian Optimization (PBO)

Preferential Bayesian Optimization [Gonzalez et al 2017] assumes a latent preference function $g : \mathcal{X} \to \mathcal{R}$ and constructs a joint reward $f([x, x']) = g(x') - g(x)$ that defines a preference function

$$\pi_f([x, x']) = \sigma(f[x, x'])$$

With an initial dataset $\mathcal{D}_0 = \{[x_i, x'_i, y_i]\}_{i=1}^N$

1. Fit a GP with kernel $k$ to $\mathcal{D}_j$ and learn $\pi_{f,j}(x)$

2. Compute posterior and estimate next pair of interests using duel-Thompson sampling AF

$$[x_{j+1}, x'_{j+1}] = \arg\max_{x \in \mathcal{X}} \int_{\mathcal{X}} \pi_f([x, x']; D_j)dx', \arg\max_{x'_* \in \mathcal{X}} \mathbb{V}[\sigma(f_*)|[x_*, x'_*], D_j, x_* = x_{j+1}]]$$

3. Augment data $\mathcal{D}_{j+1} = \{\mathcal{D}_j, ([x_{j+1}, x'_{j+1}], y_{j+1})\}$

Iterate 1 to 3 and report Condorcet's winner (who wins a majority vote in every head-to-head election against each other candidates)
Exploratory Experiments

Field study (N=2)
3 months of usage, male technical artists, age range 25-35, both 3 more years experience
Collected 549 sequence, mean iteration 4.1, range 1-23

Lab study (N=20)
each participant for 90 minutes, 7 female, no diverse, age range 18-62, 4 with 1-year experience
Collected 200 sequences, mean iteration 5.1, range 1-11

Models Used in Lab Study

Models Appeared in Field Study
Hypothesis and Intuition

If the Bayesian optimizer successfully optimized the outcome, the overall ratings in the diagram should move to the right (Higher ratings).
Observations

Large mismatch between expected and actual ratings
Either non-stationary and decreasing
Did the Optimization Work?

Yes, objectively highly simplified models were rated higher

However, subjective satisfaction does not support it: Field 11.9%, Lab 48.5%
Pitfalls: Human Side

Heuristic bias
“This is similar to examples I have been dealing with…” (But actually quite different)

Loss aversion
“I’ve seen better results before, but the results are getting worse and worse”

Diminishing returns
“I can’t see differences anymore”

Rapid Adaptation
“I changed my mind”
Pitfalls: Machine Side

Underlying algorithms often assume:

**Stable (latent) preference assumption**
“I’ve changed my mind”
“X is better in A, B, C but Y is better in D, E, F”

**Complete preference assumption**
“I don’t know”
Countermeasures

Reducing decision noise [Kahneman et al. 2021] regarding level, stable pattern and transient decision noise

**Reduce decision variability**
e.g. provide a timeline to support recall mitigating loss aversion

**Reduce contextual bias**
e.g. indicate optimization intention to frame the current context better, mitigating representativeness and availability bias and avoid judging based on previous examples

**Reduce purely occasional flaws**
e.g. present previous results to add consistency check
Empirical Study II
Rethinking Opinion Measurement Interfaces [Ou et al. ToCHI’ 2?]

![Diagram showing the interaction between user inputs, behaviors, and perceptions.](image)

**Human Intelligence**

**user interface space**

**Inputs**

**Behaviors**

**Perceives**

---

**Fig. 1.** Different opinion measurement interfaces. (UI) Single-item Rating Scale (SIRS), (II) Ternary Alternative Forced Choice (3-AFC), (III) Alternate-Free Rating (AFR), (IV) Ternary Alternatives Forced Ranking (AFR), (V) Alternate-Free Forced Ranking with Distance (AFRF-D). The SIRS and 3-AFC interfaces are rating-scale based measurements, and the 2-AFCs, U, and AFRF are preference-based ranking interfaces. A combination of the 2-AFCs relies absolute ratings and preferential ranking, which permits capturing not only ranking orders but also local ordinal ranking distances. All interfaces in this gallery also permit users to express their interrater preferences through "I don’t know," and to specify when a satisfactory result is achieved through "It’s satisfied".

Humans in the loop optimization refers to systems that present alternative choices to a human decision maker and query their judgment to optimize system parameters iteratively. The user interface typically provides either an absolute rating scale or a preferential option. However, these approaches potentially suffer from collinearity problems, incoherencies, and incompleteness violations of norm preferences in such a sequential evolution. This paper rectifies the design scope of opinion measurement interfaces, specifically for the human-in-the-loop optimization context. We conducted a study (N = 36) using a two-questionnaire task and an ergo color enhancement task to evaluate human-salient feedback in six representative opinion measurement interfaces. Based on our analysis, we recommend using a hierarchy approach to support sequential opinion measurement, which counterweights individual rating and preferential choice limitations and carries more information for subsequent stages. We further discuss the trade-offs in different user interface designs and their implications for real-world applications.

---

**Authors’ addresses:** ChangLuo, U. Munich, Germany. changluo@informatik.uni-muenchen.de, Jevan Matar, LMU Munich, Germany. ji.matar@tum.de. Daniel Büscher, University of Bayreuth, Germany. daniel.buescher@uni-bayreuth.de. Andreas Butz, LMU Munich, Germany. Andreas.Butz@Informatik.Uni-Muenchen.de.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1550-8719/20/01...
https://doi.org/10.1145/XXXXXXX.XXXXXXX

Opinion Measurement Interfaces

Amazon By Feature Ratings

- By feature
  - Value for money: ★★★★★☆☆☆☆☆ 4.2
  - Easy to install: ★★★★★☆☆☆☆☆ 4.1
  - Smart Features: ★★★★★☆☆☆☆☆ 3.7
  - Tech Support: ★★★★★☆☆☆☆☆ 3.4

Two football teams bet

- V.S.

Two candidates voting

- Candidate A
- Candidate B
- Abstain

a) Pointwise examples

Spotify’s Default Music Order

<table>
<thead>
<tr>
<th>#</th>
<th>TITLE</th>
<th>ALBUM</th>
<th>DATE ADDED</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hooked On A Feeling</td>
<td>Blue Swede, Einst Shifting</td>
<td>May 23, 2017</td>
<td>2:39</td>
</tr>
<tr>
<td>2</td>
<td>On All The Way</td>
<td>Raspberries</td>
<td>May 23, 2017</td>
<td>2:35</td>
</tr>
<tr>
<td>3</td>
<td>Spirit In The Sky</td>
<td>Music From The Motion Picture M...</td>
<td>May 23, 2017</td>
<td>4:30</td>
</tr>
<tr>
<td>4</td>
<td>Mounga Deydman - 2002 Remastered</td>
<td>David Ball</td>
<td>May 22, 2017</td>
<td>4:45</td>
</tr>
</tbody>
</table>

b) Pairwise examples

Google Scholar’s HCI Conference Ranking

<table>
<thead>
<tr>
<th>Publication</th>
<th>h5-index</th>
<th>h5-median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Computer Human Interaction (CHI)</td>
<td>147</td>
<td>154</td>
</tr>
<tr>
<td>2. IEEE Transactions on Interactive, Mobile, Wearable and Ubiquitous Technologies</td>
<td>120</td>
<td>106</td>
</tr>
<tr>
<td>3. Proceedings of the ACM on Human-Computer Interaction</td>
<td>91</td>
<td>77</td>
</tr>
<tr>
<td>4. Proceedings of the ACM on Human-Computer Interaction</td>
<td>97</td>
<td>72</td>
</tr>
<tr>
<td>5. International Journal of Human-Computer Studies</td>
<td>97</td>
<td>86</td>
</tr>
<tr>
<td>6. ACM/IEEE International Conference on Human-Computer Interaction</td>
<td>92</td>
<td>77</td>
</tr>
<tr>
<td>7. ACM Conference on Computer-Supported Cooperative Work &amp; Social Computing</td>
<td>99</td>
<td>74</td>
</tr>
</tbody>
</table>

IMDB Top 5 Movies

- The Godfather (1972) ★★★★★
- The Dark Knight (2008) ★★★★★
- The Godfather Part II (1974) ★★★★★
- The Godfather Part III (1990) ★★★★★
- The Godfather Part IV (1997) ★★★★★

Imdb Rating

- ★★★★★

47
Opinion Measurement Interfaces for HITL Systems

[Marks et al. 1997] [Wilber et al. 2014] [Brochu et al. 2007] [Welleck et al. 2020]

[Christiano et al. 2017] [Brochu et al. 2007] [Koyama et al. 2014] [Koyama et al. 2020]
Opinion Measurement UIs for HITL Systems

1-RS, n-RS: single (or n) rating scale(s), bipolar

2-ANFC: two alternative non-forced choice

n-AFR, n-ANFR: n-alternative (non-)forced ranking

n-ANFRD: n-alternative non-forced ranking with distance
Opinion Measurement UIs for HITL Systems

1-RS, n-RS: single (or n) rating scale(s), bipolar

2-ANFC: two alternative non-forced choice

n-AFR, n-ANFR: n-alternative (non-)forced ranking

n-ANFRD: n-alternative non-forced ranking with distance

<table>
<thead>
<tr>
<th>Feedback Type</th>
<th>Alternative Options (n = 1, 2, 3, …)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Utility</strong></td>
<td>1-RS (UI1)</td>
</tr>
<tr>
<td></td>
<td>2-RS</td>
</tr>
<tr>
<td></td>
<td>n-RS (UI3)</td>
</tr>
<tr>
<td><strong>Preferential (strict)</strong></td>
<td>2-AFC</td>
</tr>
<tr>
<td></td>
<td>n-AFR (UI4)</td>
</tr>
<tr>
<td><strong>Preferential (weak)</strong></td>
<td>2-ANFC (UI2)</td>
</tr>
<tr>
<td></td>
<td>n-ANFR (UI5)</td>
</tr>
<tr>
<td><strong>Hybrid (strict)</strong></td>
<td>2-ANFC (UI2)</td>
</tr>
<tr>
<td></td>
<td>n-ANFRD</td>
</tr>
<tr>
<td><strong>Hybrid (weak)</strong></td>
<td>n-ANFRD</td>
</tr>
</tbody>
</table>

Hybrid

<table>
<thead>
<tr>
<th>Feedback Type</th>
<th>Alternative Options (n = 1, 2, 3, …)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Utility</strong></td>
<td>1-RS (UI1)</td>
</tr>
<tr>
<td></td>
<td>2-RS</td>
</tr>
<tr>
<td></td>
<td>n-RS (UI3)</td>
</tr>
<tr>
<td><strong>Preferential (strict)</strong></td>
<td>2-AFC</td>
</tr>
<tr>
<td></td>
<td>n-AFR (UI4)</td>
</tr>
<tr>
<td><strong>Preferential (weak)</strong></td>
<td>2-ANFC (UI2)</td>
</tr>
<tr>
<td></td>
<td>n-ANFR (UI5)</td>
</tr>
<tr>
<td><strong>Hybrid (strict)</strong></td>
<td>2-ANFC (UI2)</td>
</tr>
<tr>
<td></td>
<td>n-ANFRD</td>
</tr>
<tr>
<td><strong>Hybrid (weak)</strong></td>
<td>n-ANFRD</td>
</tr>
</tbody>
</table>
Opinion Measurement UIs for HITL Systems

1-RS, n-RS: single (or n) rating scale(s), bipolar

2-ANFC: two alternative non-forced choice

n-AFR, n-ANFR: n-alternative (non-)forced ranking

n-ANFRD: n-alternative non-forced ranking with distance

<table>
<thead>
<tr>
<th>Feedback Type</th>
<th>Pointwise (n=1)</th>
<th>Pairwise (n=2)</th>
<th>Listwise (n&gt;2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility</td>
<td>1-RS (UI1)</td>
<td>2-RS</td>
<td>n-RS (UI3)</td>
</tr>
<tr>
<td>Preferential (strict)</td>
<td>2-AFC</td>
<td>n-AFR (UI4)</td>
<td></td>
</tr>
<tr>
<td>Preferential (weak)</td>
<td>2-ANFC (UI2)</td>
<td>n-ANFR (UI5)</td>
<td></td>
</tr>
<tr>
<td>Hybrid (strict)</td>
<td></td>
<td>n-AFRD</td>
<td></td>
</tr>
<tr>
<td>Hybrid (weak)</td>
<td></td>
<td>n-ANFRD (UI6)</td>
<td></td>
</tr>
</tbody>
</table>

Alternative Options (n = 1, 2, 3, …)
Opinion Measurement UIs for HITL Systems

1-RS, n-RS: single (or n) rating scale(s), bipolar

2-ANFC: two alternative non-forced choice

n-AFR, n-ANFR: n-alternative (non-)forced ranking

n-ANFRD: n-alternative non-forced ranking with distance

<table>
<thead>
<tr>
<th>Feedback Type</th>
<th>Alternative Options (n = 1, 2, 3, …)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility</td>
<td>1-RS (UI1)</td>
</tr>
<tr>
<td></td>
<td>2-RS</td>
</tr>
<tr>
<td></td>
<td>n-RS (UI3)</td>
</tr>
<tr>
<td>Preferential (strict)</td>
<td>2-AFC</td>
</tr>
<tr>
<td></td>
<td>n-AFR (UI4)</td>
</tr>
<tr>
<td>Preferential (weak)</td>
<td>2-ANFC (UI2)</td>
</tr>
<tr>
<td></td>
<td>n-ANFR (UI5)</td>
</tr>
<tr>
<td>Hybrid (strict)</td>
<td>n-ANFRD (UI6)</td>
</tr>
<tr>
<td>Hybrid (weak)</td>
<td>n-ANFRD (UI6)</td>
</tr>
</tbody>
</table>

Available UIs:

- UI1 (1-RS)
- UI2 (2-ANFC)
- UI3 (n-RS)
- UI4 (n-AFR)
- UI5 (n-ANFR)
- UI6 (n-ANFRD)
Hypotheses for UI Ranking

H1 Baseline

- UI1 < UI2
Hypotheses for UI Ranking

H1 Baseline
- UI1 < UI2

H2 Impact of Listwise UI
- UI1 < UI3
- UI2 < UI5
Hypotheses for UI Ranking

H1 Baseline
- UI1 < UI2

H2 Impact of Listwise UI
- UI1 < UI3
- UI2 < UI5

H3 Impact of Listwise Design Variation
- UI3 < UI4 < UI5

<table>
<thead>
<tr>
<th>Feedback Type</th>
<th>Pointwise (n=1)</th>
<th>Pairwise (n=2)</th>
<th>Listwise (n&gt;2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Utility</strong></td>
<td>1-RS (UI1)</td>
<td>2-RS</td>
<td>n-RS (UI3)</td>
</tr>
<tr>
<td>Preferential (strict)</td>
<td>2-AFC</td>
<td></td>
<td>n-AFR (UI4)</td>
</tr>
<tr>
<td>Preferential (weak)</td>
<td>2-ANFC (UI2)</td>
<td></td>
<td>n-ANFR (UI5)</td>
</tr>
<tr>
<td>Hybrid (strict)</td>
<td></td>
<td></td>
<td>n-AFRD</td>
</tr>
<tr>
<td>Hybrid (weak)</td>
<td></td>
<td></td>
<td>n-ANFRD (UI6)</td>
</tr>
</tbody>
</table>
Hypotheses for UI Ranking

H1 Baseline
- UI1 < UI2

H2 Impact of Listwise UI
- UI1 < UI3
- UI2 < UI5

H3 Impact of Listwise Design Variation
- UI3 < UI4 < UI5

H4 Impact of Hybrid UI
- UI3 < UI2 < UI4 < UI5 < UI6
User Study and Workflow

Task:
Fill beginning survey, providing feedback to the AI results; inspect improved results; loop until satisfaction, and fill ending survey
User Study and Workflow

Task:
Fill beginning survey, providing feedback to the AI results; inspect improved results; loop until satisfaction, and fill ending survey

Selection Criteria:
- A task should partially involve rational, objective judgment, and subjective components
- Each domain requires different levels of human expertise
User Study and Workflow

Task:
Fill beginning survey, providing feedback to the AI results; inspect improved results; loop until satisfaction, and fill ending survey

Selection Criteria:
- A task should partially involve rational, objective judgment, and subjective components
- Each domain requires different levels of human expertise

Participants (N=2x6x30=360)
2 selected domains. 30 for each UI, 171 female, 185 male, 4 diverse; age range 18-66 (M=28.14)
Apparatus: Text Summarization

Pre-trained BART model [Lewis et al. 2019] fine-tuned for CNN, nucleus sampling

4 adjustable hyper parameters
- Summarization ratio
- Length penalty
- top-p
- Temperature
Apparatus: Photo Color Enhancement


4 adjustable hyper parameters
- Brightness
- Contrast
- Saturation
- Color Temperature
**Apparatus: Bayesian Optimizer**

Based on Expected Utility Bayesian Optimization [Lin et al. 2022]

Modified to fit objects ranking optimization

| Example outcomes |
|------------------|------------------|------------------|------------------|
| Barcelona beat Atletico Madrid 3-0 to remain in touch with Real Madrid in La Liga. Lionel Messi and Deigo scored for Barca in Barca’s fourth straight league win against hi-g | Barcelona beat Atletico Madrid 3-0 to stay in touch with Primera Liga leaders Real Madrid. Real beat Recreativo Huelva 2-0 and Real Madrid beat Real 2-1 to stay two points clear of Real. Real’s first-half goal scorer Galo Higavain 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 Huelva 2-0 and Real Madrid beat Real 2-1 to stay two points clear of Real. Real’s first-half goal scorer Galo Higavain 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 hi-g | 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 hi-g | 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 6th straight league game. Real Madrid beat Recreativo Huelva 2-0 and Galo Higavain scored in the dying minutes. Real have made their best start since 1993 but coach Bernd Schuster's rotation policy questioned. |
Recall: Preferential Bayesian Optimization

Preferential Bayesian Optimization [Gonzalez et al 2017] assumes a latent preference function \( g : \mathcal{X} \rightarrow \mathcal{R} \) and constructs a joint reward \( f([x, x']) = g(x') - g(x) \) that defines a preference function

\[
\pi_f([x, x']) = \sigma(f[x, x'])
\]

With an initial dataset \( D_0 = \{(x_i, x'_i, y_i)\}_{i=1}^N \)

1. Fit a GP with kernel \( k \) to \( D_j \) and learn \( \pi_{f, j}(x) \)

2. Compute posterior and estimate next pair of interests using duel-Thompson sampling

\[
[x_{j+1}, x'_{j+1}] = \operatorname{argmax}_{x \in \mathcal{X}} \int \pi_f([x, x']; D_j) dx', \operatorname{argmax}_{x' \in \mathcal{X}} \mathbb{V}[\sigma(f_*)|[x_*, x'_*], D_j, x_* = x_{j+1}]
\]

3. Augment data \( D_{j+1} = \{D_j, ([x_{j+1}, x'_{j+1}], y_{j+1})\} \)

Iterate 1 to 3 and report Condorcet's winner (who wins a majority vote in every head-to-head election against each other candidates)
The AF of PBO [Gonzalez et al 2017] is based on a Copeland score and the utility is unknown, which is not an absolute utility measure hence making it harder to analyze changes in the overall optimization process.

EUBO [Lin et al. 2022] considers a knowledge gradient family as AF to maximize the best options difference between iterations:

$$V(x, x') = E_j \left[ \operatorname{argmax}_{x \in \mathcal{X}} E_{j+1}[f(x)] - \operatorname{argmax}_{x \in \mathcal{X}} E_j[f(x)] \right]$$
Expected Utility Bayesian Optimization (EUBO)

The AF of PBO [Gonzalez et al. 2017] is based on a Copeland score and the utility is unknown, which is not an absolute utility measure hence making it harder to analyze changes in the overall optimization process.

EUBO [Lin et al. 2022] considers a knowledge gradient family as AF to maximize the best options difference between iterations:

$$V(x, x') = \mathbb{E}_j \left[ \arg\max_{x \in \mathcal{X}} \mathbb{E}_{j+1}[f(x)] - \arg\max_{x \in \mathcal{X}} \mathbb{E}_j[f(x)] \right]$$

Maximizing this AF is equivalent to

$$\arg\max_{x, x' \in \mathcal{X}} \text{EUBO}(x, x') = \mathbb{E}_j[\max\{f(x), f(x')\}] \subseteq \arg\max_{x, x' \in \mathcal{X}} V(x, x')$$

This provides a meaningful way to inspecting absolute latent preference utility over iterations.
HITL Optimization Performance Indicators

Machine performance
- Objective outcome quality measures
  - BLEU, ROUGE; MSE chan diff in HSV and YUV space
- Optimizer measures
  - Posterior mean of the estimated ranking utility (*Latent Utility*)

User performance
- Interaction behavior measures
  - decision time, iterations, incomplete/indifference preference, ranking interactions
- User’s actual input rating/ranking utility for the machine outcomes (*Direct Utility*)
- subjective satisfaction
HITL Optimization Performance Indicators

Machine performance
- Objective outcome quality measures
  - BLEU, ROUGE; MSE chan diff in HSV and YUV space
- Optimizer measures
  - Posterior mean of the estimated ranking utility (Latent Utility)

User performance
- Interaction behavior measures
  - decision time, iterations, incomplete/indifference preference, ranking interactions
- User’s actual input rating/ranking utility for the machine outcomes (Direct Utility)
- subjective satisfaction

Model formula: \( \text{perf} \sim \text{UI} \times \text{progress} + (1|\text{participant}) + (1|\text{task}) \)
Barcelona's 3-0 win over Atletico Madrid. Barcelona had thumped Atletico 6-0 on 3-0 to stay in touch with Primera Liga leaders Real Madrid. Lionel Messi scores sixth goal for the Catalan giants. Real Madrid beat Recreativo Huleva 2-0 in their La Liga clash to stay two points clear. Real's first-ever league win since 1991 as Real Madrid make their best start since 1991. Real's Abel Resino becomes first Spanish league coach to be sacked this season.

Objective 1
Provide feedback using the ranking interface about summarized text to archive these objectives:

- Strongly Agree
- Agree
- I don't know
- Disagree
- Strongly Disagree

I'M SATISFIED
SUBMIT FEEDBACK
User Performance: Interaction Behavior

Users spend more time when using listwise interfaces (UI3, UI4, UI5, UI6)

Text Summarization Task

Photo Color Enhancement Task
User Performance: Interaction Behavior

Users spend more time when using listwise interfaces (UI3, UI4, UI5, UI6)

Participants express more incomplete and indifference judgements (UI2, UI3, UI4, UI5, UI6)
Machine Performance: Utility (Direct and Latent)

UI6 significantly outperforms other UIs in terms of direct preference utility.
UI6 performed similarly compared to UI2 in terms of latent preference utility inferred by optimizer.

Machine Performance: Utility (Direct and Latent)
Machine Performance: Outcome (Objective) Quality

UI6 outperforms other UIs in terms of performance in various aspects

Text Summarization Task

Photo Color Enhancement Task
Opinion Measurement UI Performance

UI Ranking Hypothesis
Opinion Measurement UI Performance

UI Ranking Hypothesis

UI1  UI3  UI2  UI4  UI5  UI6

Decision Time

Human Feedback

UI1  UI2  UI3  UI4  UI5  UI6
Optimal Opinion Measurement UIs for HITL

- Pointwise UI (UI1) should be avoided in general.
- Pairwise UIs (e.g., UI2) favor fast decisions but are limited by consistent ranking order assumptions.
- Listwise UIs without specified ranking distance (e.g., UI3, UI4, UI5) do not have general advantages over pairwise approach, but permitting a ranking distance (e.g., UI6) collects more information from users, and favors better optimized results.
- The choice of n-ANFRD (UI6) or 2-ANFC (UI2) is a decision tradeoff between decision time and overall optimization performance.
Empirical Study III
Expertise Considered Harmful? [Ou et al. IUI’ 23]

The Impact of Expertise in the Loop for Exploring Machine Rationality
Changliun Ou
LAI Branch
Germany
research@changliun.de

Sven Mayer
LAI Branch
Germany
infotech-mayer.com

Andreas Itz
LAI Branch
Germany
basti.itz.de

Figure 6: The interface for optimizing AI model simplification using an expert in the loop. In each iteration, the interface presents four 3D models. Participants can drag and drop the top right models to a variable rating region to indicate their rating or submission. Each of the regions can contain multiple models. Models can be set to “I don’t know” to express an inconspicuous performance. Participants can indicate their satisfaction by increasing the optimization bar. To inspect the AI model quality, they can move the bar up or down and rotate the 3D models simultaneously using a mouse.

ABSTRACT
Humans in the loop optimization leverage human expertise to guide machine operations iteratively and search for an optimal solution to a solution space. While prior empirical studies mostly investigated various AI models, we focused on the impact of the levels of expertise on the vector’s quality and corresponding subjective evaluation. We conducted a study (N=24) in virtual reality, and 3D model simplification using an expert in the loop for optimization. Participants with higher expertise led to more optimization iterations but with more explicit performance while keeping satisfaction. In contrast, novice users were more satisfied and introduced fewer iterations. Therefore, we conclude that experts with user interface experience with the machine models optimal models, and the observed behavior can be explored as a performance indicator for the loop system design to improve matching models. We include future research in the conclusions about the impact on more expertise when designing human in the loop systems.

CCS CONCEPTS
• Human-centered computing → Empirical studies in HCC
• Interaction paradigms → Computing methodologies → Active learning styles.

KEYWORDS
human in the loop machine learning, adaptive human computer interaction, reliability.

ACM Reference Format:
Observation and Hypothesis

Observations:
1. Context-dependent, unstable preference objectives can lead to non-satisfactory results
2. Prior HITL optimization systems are mostly evaluated on novice users and rarely report on the effect of user expertise

Hypothesis:
Using higher expertise leads to better results in HITL optimization
User Study and Workflow

Task:
Fill beginning survey, providing feedback to the AI results; inspect improved results; loop until satisfaction, and fill ending survey

Selection Criteria:
- A task should partially involve rational, objective judgment, and subjective components.
- Each domain requires different levels of human expertise

Participants (N=60)
3 selected domains. 20 for each domain, 31 female, 29 male; age range 19-52 (M=26.92)
Apparatus

Pre-trained BART model [Lewis et al. 2019] fine-tuned for CNN, nucleus sampling


A parameterized 3D mesh simplifier [Ou et al. 2022]
Apparatus: 3D Model Simplification

A parameterized 3D mesh simplifier [Ou et al. 2022]

5 adjustable hyper parameters

- Simplification ratio
- Border preservation
- Hard edge preservation
- Sharpness preservation
- Quadrilateral preservation
HITL Optimization Performance Indicators

Machine performance
- Objective outcome quality measures
  - BLEU, ROUGE; MSE chan diff in HSV and YUV space; SSIM/PSNR, Jacobian Cell, Chamfer
- Optimizer measures
  - Posterior mean of the estimated ranking utility [Chu and Grahamani, 2005] (Latent Utility)

User performance
- Interaction behavior measures
  - decision time, iterations, incomplete/indifference preference, ranking interactions
- User’s actual ranking utility for the machine outcomes (Direct Utility)
- subjective satisfaction, expertise measures (e.g. years of expertise)
  - quantile-based discretization (derive relative expertise)
HITL Optimization Performance Indicators

Machine performance

- Objective outcome quality measures
  - BLEU, ROUGE; MSE chan diff in HSV and YUV space; SSIM/PSNR, Jacobian Cell, Chamfer
- Optimizer measures
  - Posterior mean of the estimated ranking utility [Chu and Grahamani, 2005] (Latent Utility)

User performance

- Interaction behavior measures
  - decision time, iterations, incomplete/indifference preference, ranking interactions
- User’s actual ranking utility for the machine outcomes (Direct Utility)
- subjective satisfaction, expertise measures (e.g. years of expertise)
  - quantile-based discretization (derive relative expertise)

Model formula: \( \text{perf} \sim \text{expertiseLevel} \times \text{progress} + (1|\text{participant}) + (1|\text{task}) \)
Machine Performance: Objective Outcome Quality

Novices and intermediates can reach expert level performance
User Performance: Interaction Behavior

Experienced participants explore solutions more when feedback loop is more efficient
User Performance: Interaction Behavior

Experienced participants explore solutions more when feedback loop is more efficient.

Experienced participants indicate a clearer preference by showing more frequently incomplete/indifferent preferences.
Instead, **novices** are significantly **more satisfied** than experienced ones.
Solution Space

Parameter A

Parameter B
Performance Space

Metric B

Metric A
Performance Space

Metric B

Optimizer

Metric A
Performance Space

\[ M_A(\text{red rabbit}) \geq M_A(\text{blue rabbit}) \]

\[ M_B(\text{red rabbit}) \geq M_B(\text{blue rabbit}) \]

\[ \Rightarrow \quad \text{dominants} \]
Performance Space

Metric B

Optimizer

Metric A
Performance Space

Pareto Front
[Pareto, 1912]

Metric A

Metric B

Optimizer
Implications on the Impact of Expertise in HITL

- Novice
- Intermediate
- Experienced
- Expert

Pareto Front

Final Outcome Quality

Involved Expertise

Quality
Implications on the Impact of Expertise in HITL

Pareto Front

Random Feedback Generator

Novice

Intermediate

Experienced

Expert

Final Outcome Quality

Involved Expertise

?
Implications on the Impact of Expertise in HITL

The Borel–Cantelli lemma [Borel 1909] [Cantelli 1917]
With infinite amount of events, the probability of observing any meaningful result is 1.0

$$\sum_{n=1}^{\infty} P(E_n) = +\infty \Rightarrow P(\{E_n \text{ i.o.}\}) = 1$$
Implications on the Impact of Expertise in HITL

- Pareto Front
- Novice
- Intermediate
- Experienced
- Expert
- Time Needed To Find Optimal
- Final Outcome Quality
- Random Feedback Generator
- Borel–Cantelli
- Involved Expertise
Implications on the Impact of Expertise in HITL

- Pareto Front
- Random Feedback Generator
- Novice, Intermediate, Experienced, Expert
- Borel–Cantelli
- Involved Expertise
- Time Needed to Find Optimal
- Final Outcome Quality
- Subjective Satisfaction
Implications on the Impact of Expertise in HITL

Subjective Satisfaction vs. Time Needed to Find Optimal

Random Feedback Generator vs. Novice, Intermediate, Experienced, Expert

Borel–Cantelli Implications on the Impact of Expertise in HITL
Reflections
Building Blocks in HITL Optimization Systems

HITL Optimization System

Objective

World

Circle: problem context

Arrow: behaviors

Circle: Objective World

User

Inputs
Building Blocks in HITL Optimization Systems

- Human Intelligence
- Machine Intelligence
- Objective World

User

Inputs

Behaviors

Problem Context
Building Blocks in HITL Optimization Systems

Empirical Knowledge

inputs
behaviors

user

user interface space

problem context

Human Intelligence

Machine Intelligence

Objective World
Building Blocks in HITL Optimization Systems

User interface space

Empirical Knowledge

performance space

inputs

measures

Human Intelligence

Machine Intelligence

Objective World

problem context
Building Blocks in HITL Optimization Systems
Building Blocks in HITL Optimization Systems

Objective

Empirical Knowledge

Human Intelligence

User Interface Space

User perceives inputs, which influence behaviors.

Machine Intelligence

Solution Space

Machine intelligence renders performance measures, optimizing characterizations and renders problem context.

Problem Context

Objective World
Contributions

Theoretical contribution

○ A framework that analyzes the building blocks of HITL optimization systems
○ Interpretation of when HITL is beneficial
○ Taxonomy of HITL Input UIs

Empirical and artifact contribution

○ All work are open sourced
○ https://changkun.de/s/intelligence-in-the-loop

Methodological contribution

○ A methodology for analyzing an HITL optimization system
Reflections

- Machines are designed to reproduce rational components of human intelligence.

- With more accumulated rationality in computer systems, or AI, human might only need to involve their “irrational” or subjective components that select the intended objectives based on their fundamental value or belief collected from experience.

- Using human intelligence in the loop is beneficial if and only if the machine can identify and adapt to the core value or belief of the interacting human.

- If we acknowledge the existence of our free will, we can always initiate new objectives without being considered by crowd wisdom, and will not fit into any computing systems. If we believe we can create a rational computer to cover this free will, we have to believe ourselves can be reproduced by the others.

- Creativity as part of our intelligence can be largely inspired by machine intelligence. However, the source of our individual intelligence will never be taken over by other objects.
Future Work
Unstructured and Unaggregated Feedback

- Involving unstructured and unaggregated human feedback, e.g. reinforcement learning from human feedback using proximal policy optimization [Schulman et al. 2017]
Unstructured and Unaggregated Feedback

- Involving unstructured and unaggregated human feedback, e.g. reinforcement learning from human feedback using proximal policy optimization [Schulman et al. 2017]

Prompt: “An astronaut riding a bunny in photorealistic style.”
Modeling Indifference and Incompleteness

- Involving unstructured and unaggregated human feedback, e.g. reinforcement learning from human feedback using proximal policy optimization [Schulman et al. 2017]

- Modeling indifference and incomplete preference [Nguyen et al. 2021] [Nielsen and Rigotti, 2022]
Simulating Human Prior

- Involving unstructured and unaggregated human feedback, e.g. reinforcement learning from human feedback using proximal policy optimization [Schulman et al. 2017]
- Modeling indifference and incomplete preference [Nguyen et al. 2021] [Nielsen and Rigotti, 2022]
- Simulating human priors [Murray-Smith et al. 2022]
Collective Optimization

- Involving unstructured and unaggregated human feedback, e.g. reinforcement learning from human feedback using proximal policy optimization [Schulman et al. 2017]

- Modeling indifference and incomplete preference [Nguyen et al. 2021] [Nielsen and Rigotti, 2022]

- Simulating human priors [Murray-Smith et al. 2022]

- Collective optimization [Ou et al. 2019]
Exploring Pareto Front

- Involving unstructured and unaggregated human feedback, e.g. reinforcement learning from human feedback using proximal policy optimization [Schulman et al. 2017]
- Modeling indifference and incomplete preference [Nguyen et al. 2021] [Nielsen and Rigotti, 2022]
- Simulating human priors [Murray-Smith et al. 2022]
- Collective optimization [Ou et al. 2019]
- Understanding human behaviors on a Pareto front and exploring mismatches between human and machine intelligence
Exploring Pareto Front

- Involving unstructured and unaggregated human feedback, e.g. reinforcement learning from human feedback using proximal policy optimization [Schulman et al. 2017]
- Modeling indifference and incomplete preference [Nguyen et al. 2021] [Nielsen and Rigotti, 2022]
- Simulating human priors [Murray-Smith et al. 2022]
- Collective optimization [Ou et al. 2019]
- Understanding human behaviors on a Pareto front and exploring mismatches between human and machine intelligence
Exploring Pareto Front

- Involving unstructured and unaggregated human feedback, e.g., reinforcement learning from human feedback using proximal policy optimization [Schulman et al. 2017]
- Modeling indifference and incomplete preference [Nguyen et al. 2021] [Nielsen and Rigotti, 2022]
- Simulating human priors [Murray-Smith et al. 2022]
- Collective optimization [Ou et al. 2019]
- Understanding human behaviors on a Pareto front and exploring mismatches between human and machine intelligence

[Ou et al. 202?]
Building Block in Human-in-the-loop Optimization Systems

UI

Perf.

Ctx.

Sol.

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 15.02.2023

Progressive Mesh Simplification

Expect vs. Actual Behavior

INDIFFERENCE VS. INCOMPLETENESS

BAYESIAN OPTIMIZATION

EXPLORING PARETO FRONT

COLLECTIVE

OPINION MEASUREMENT UIs

SUBJECTIVE SATISFACTION

FINAL OUTCOME QUALITY

PARETO FRONT

TIME NEEDED TO FIND OPTIMAL

INCOMPLETENESS

INSENSIBLE HUMAN EXPERTISE

PROGRESSIVE MESH SIMPLIFICATION

INITIAL VS. EXPECTED BEHAVIOR

FINAL OUTCOME QUALITY

NOVICE INTERMEDIATE EXPERIENCED EXPERT

BAYESIAN OPTIMIZATION

SUBJECTIVE SATISFACTION