LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Media Informatics Group

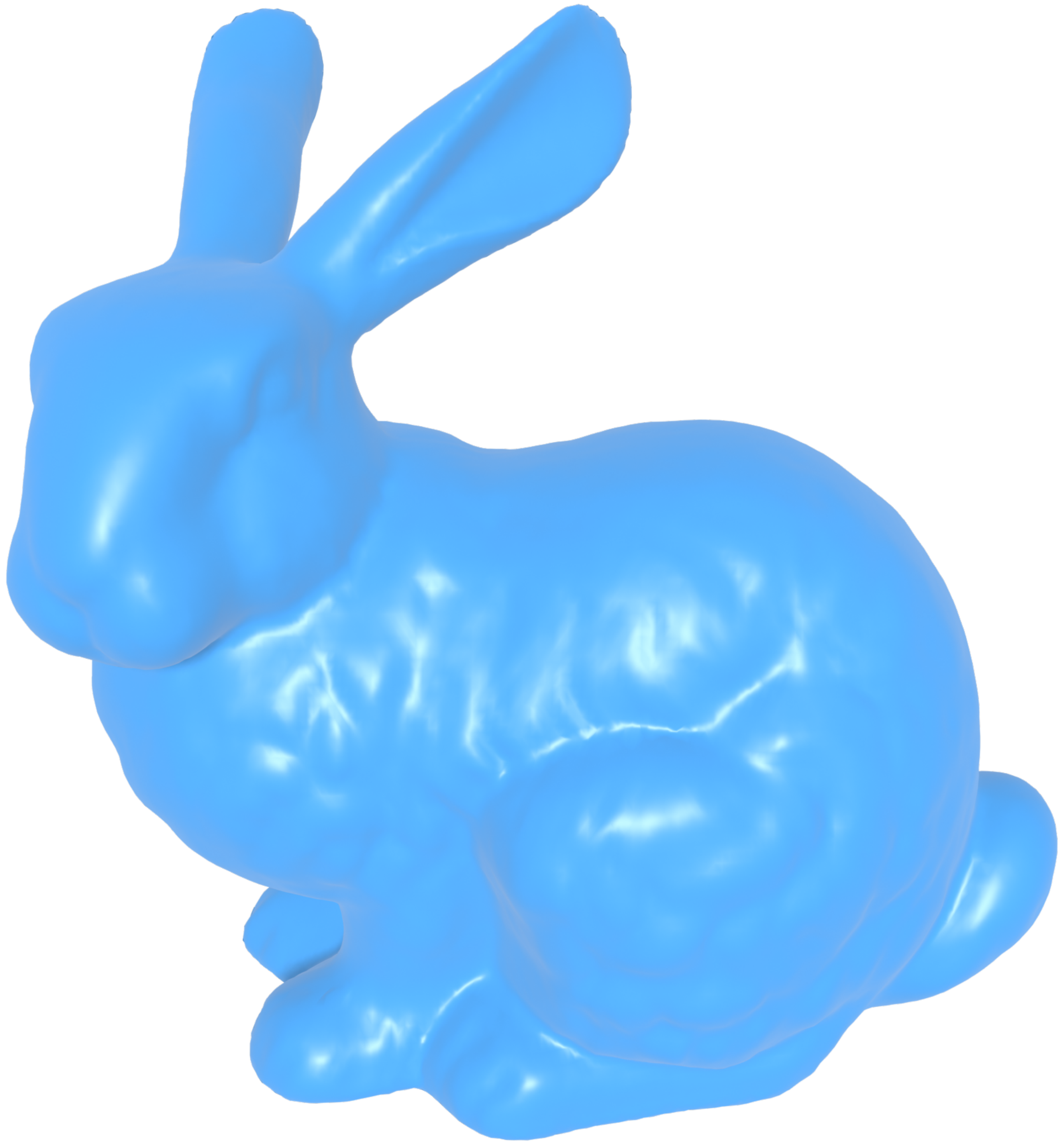
The *Intelligence* in the Loop

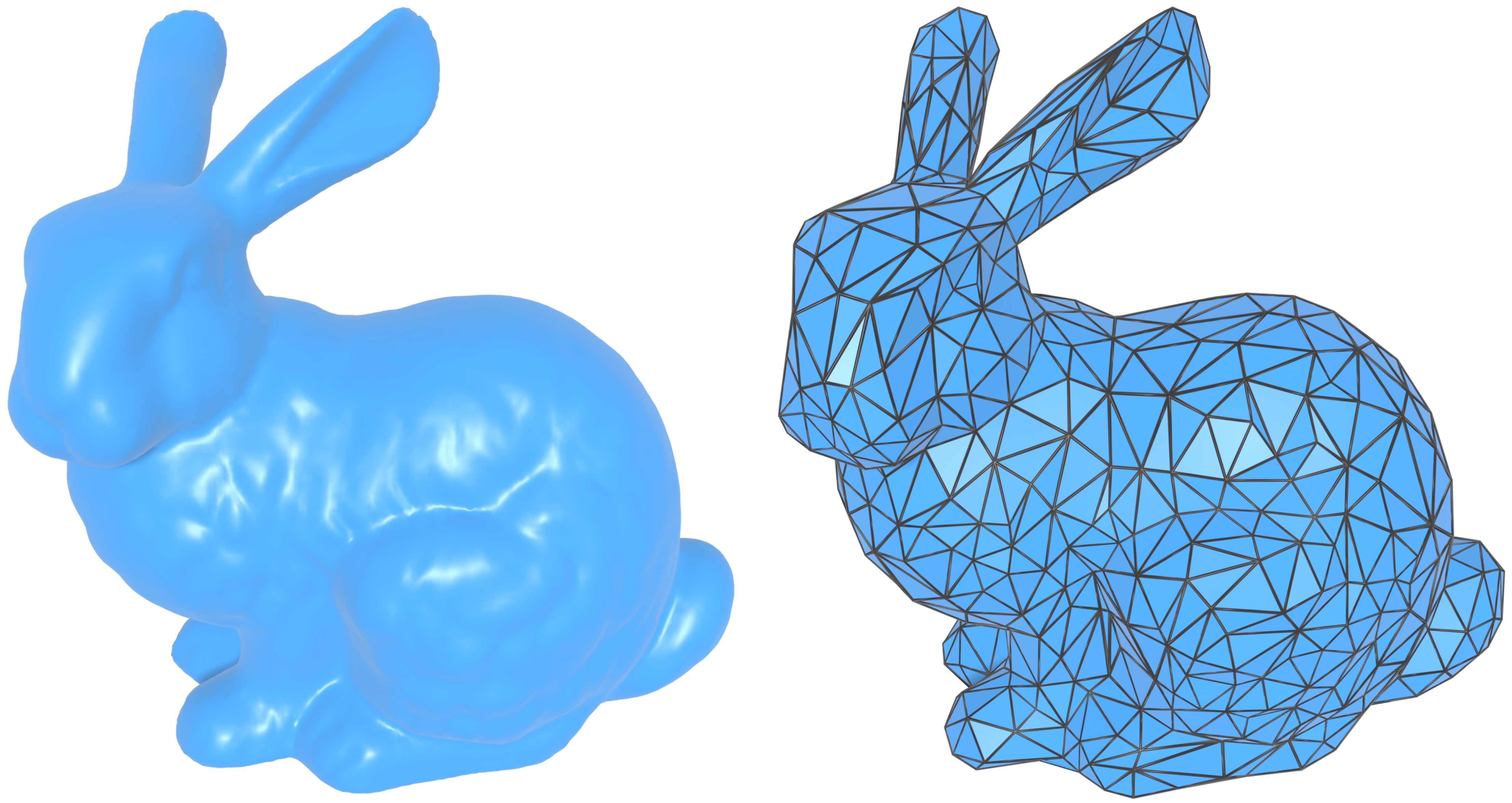
Empirical Explorations and Reflections

Changkun Ou

Ph.D. Defense | Munich, Germany | April 3rd, 2023

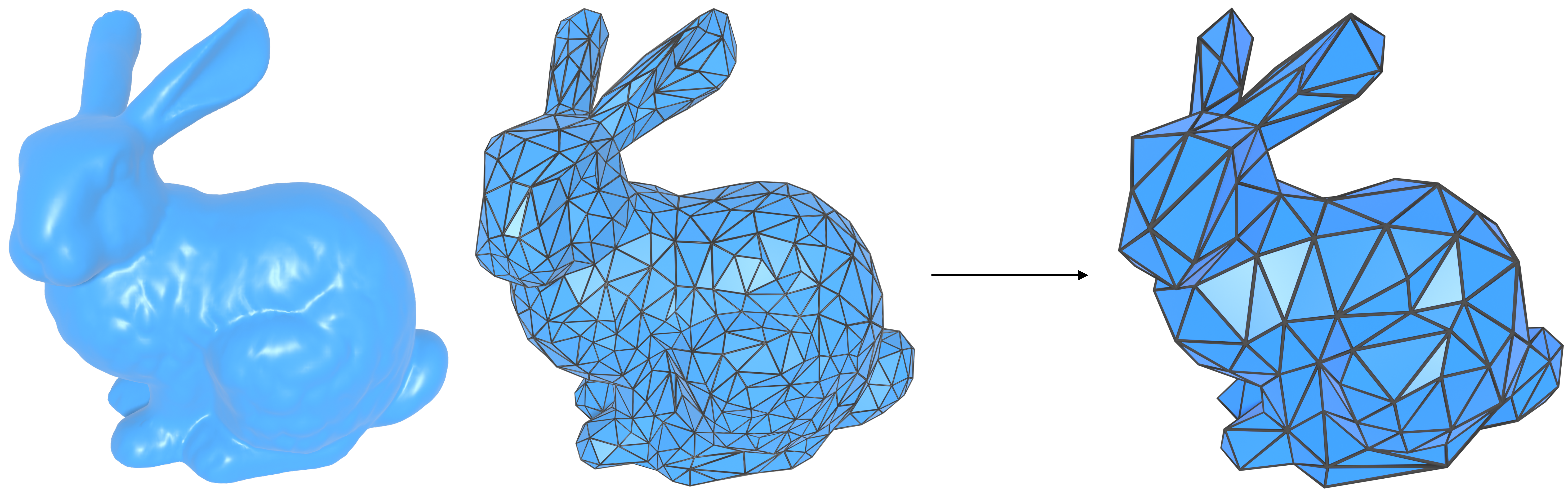
Motivation

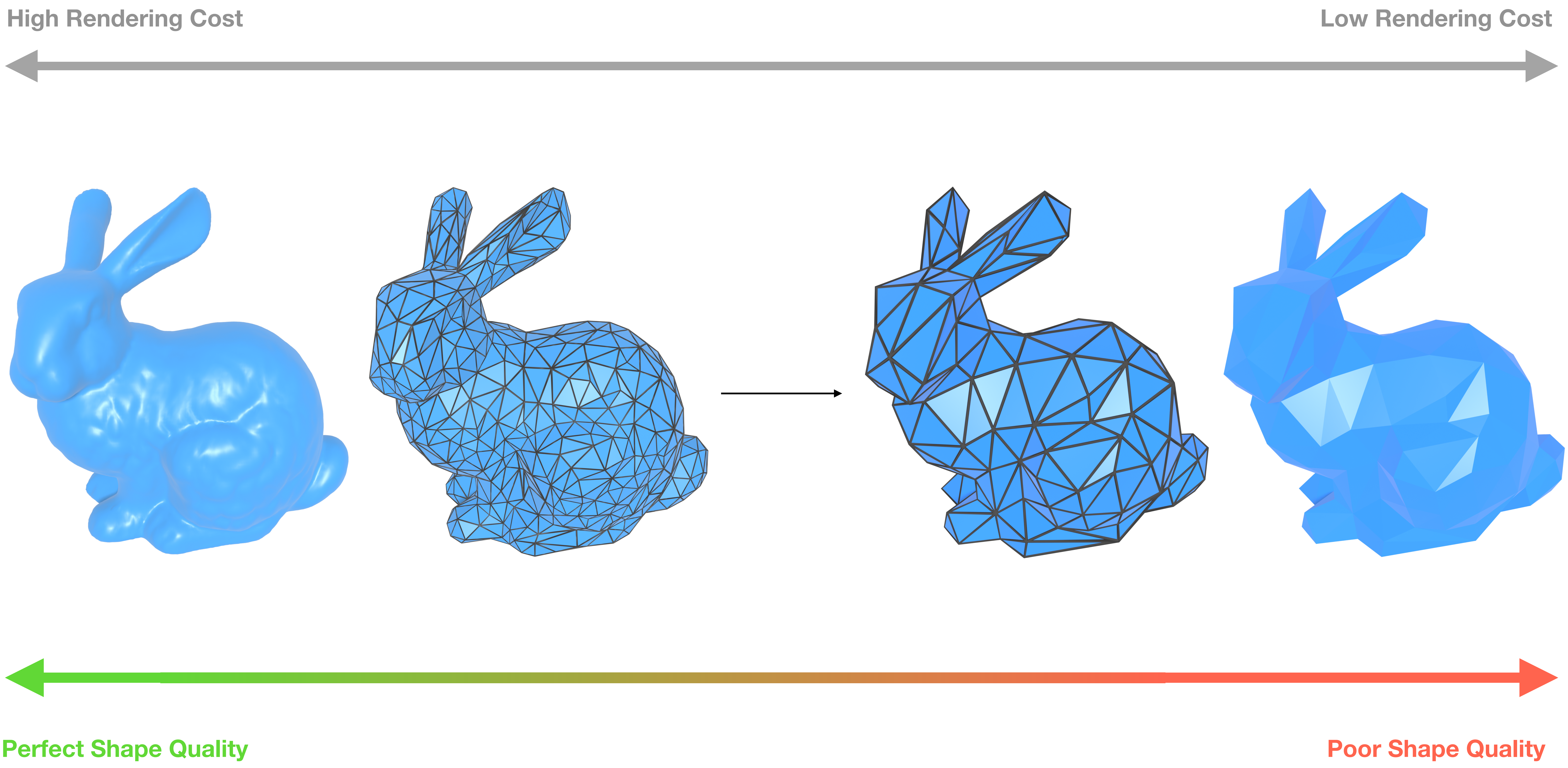




High Rendering Cost

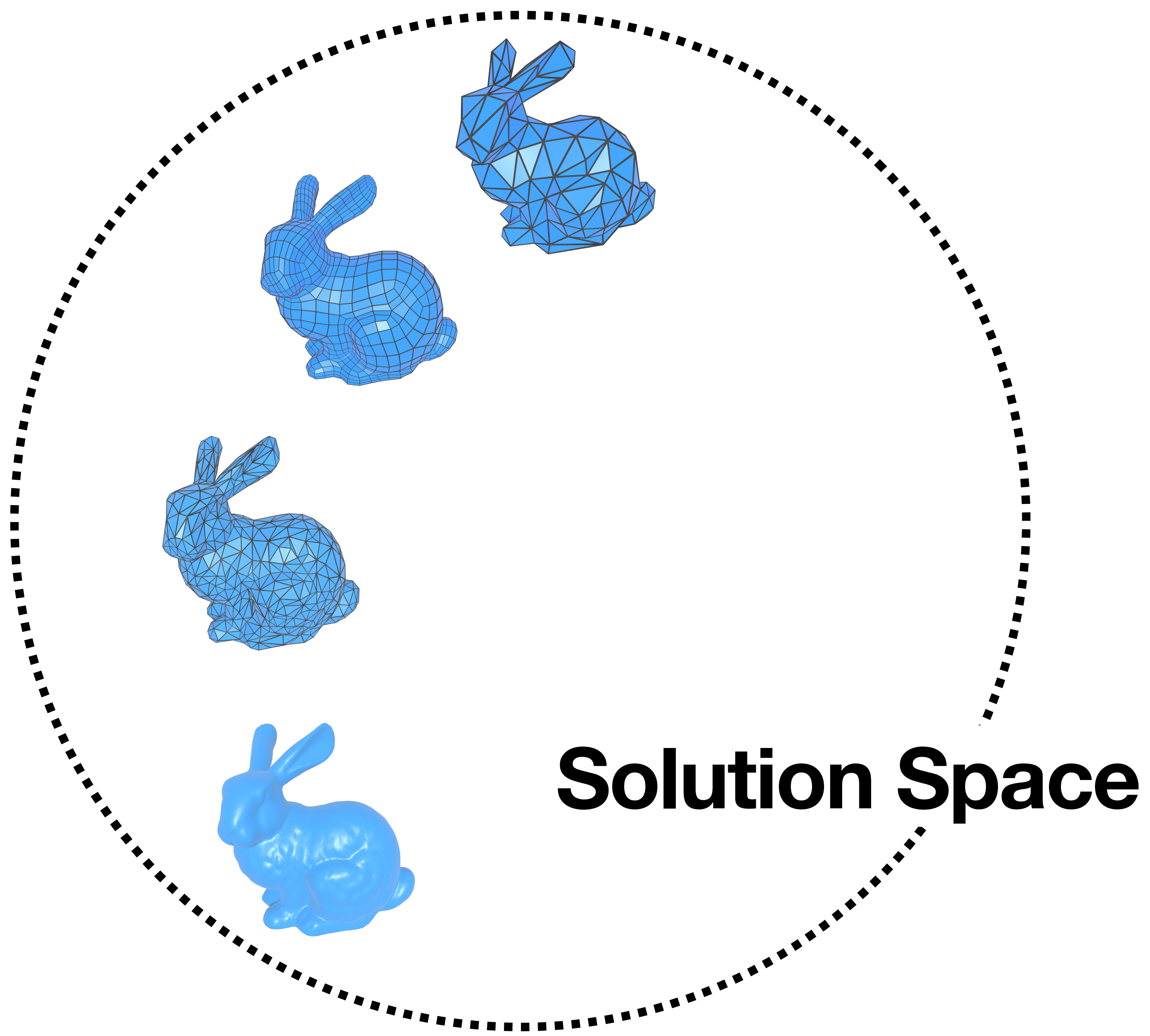
Low Rendering Cost

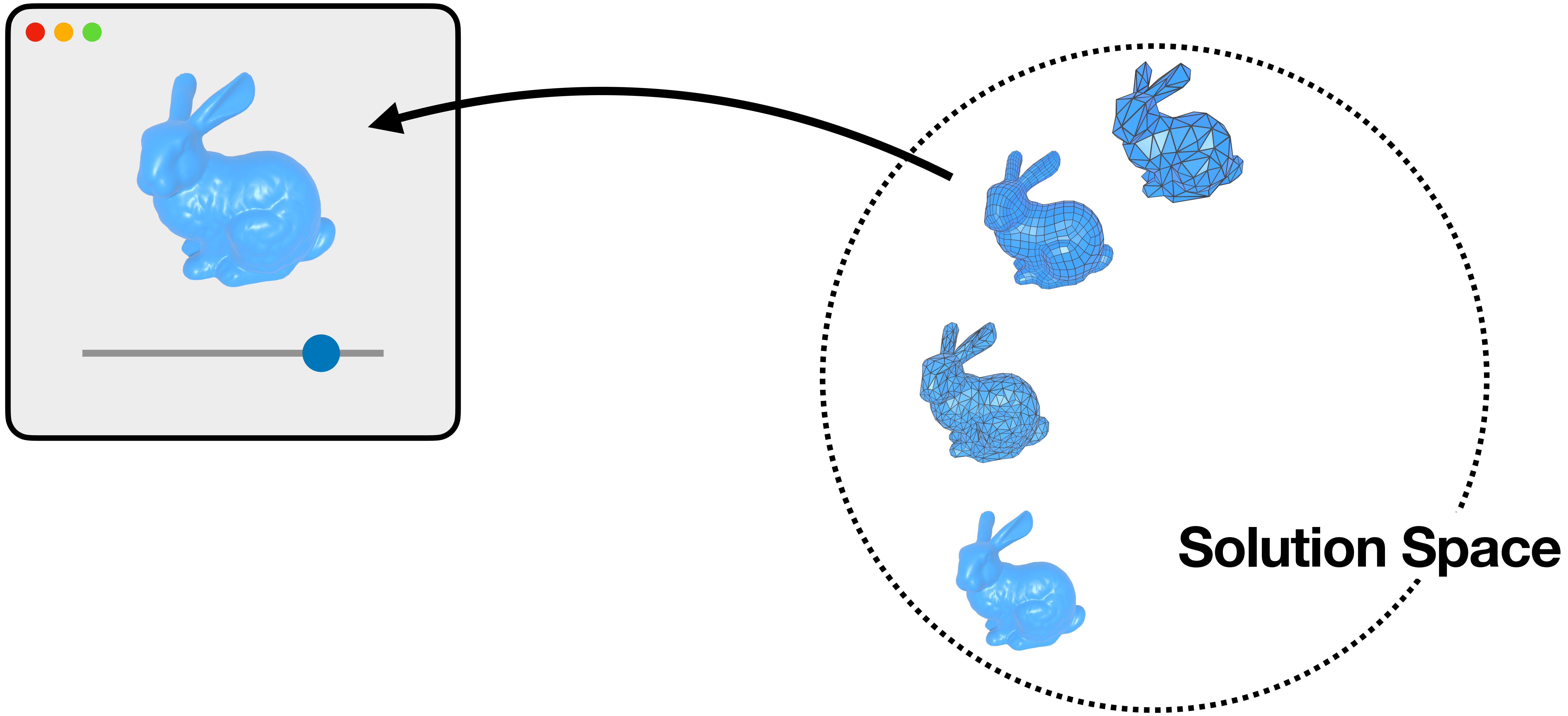


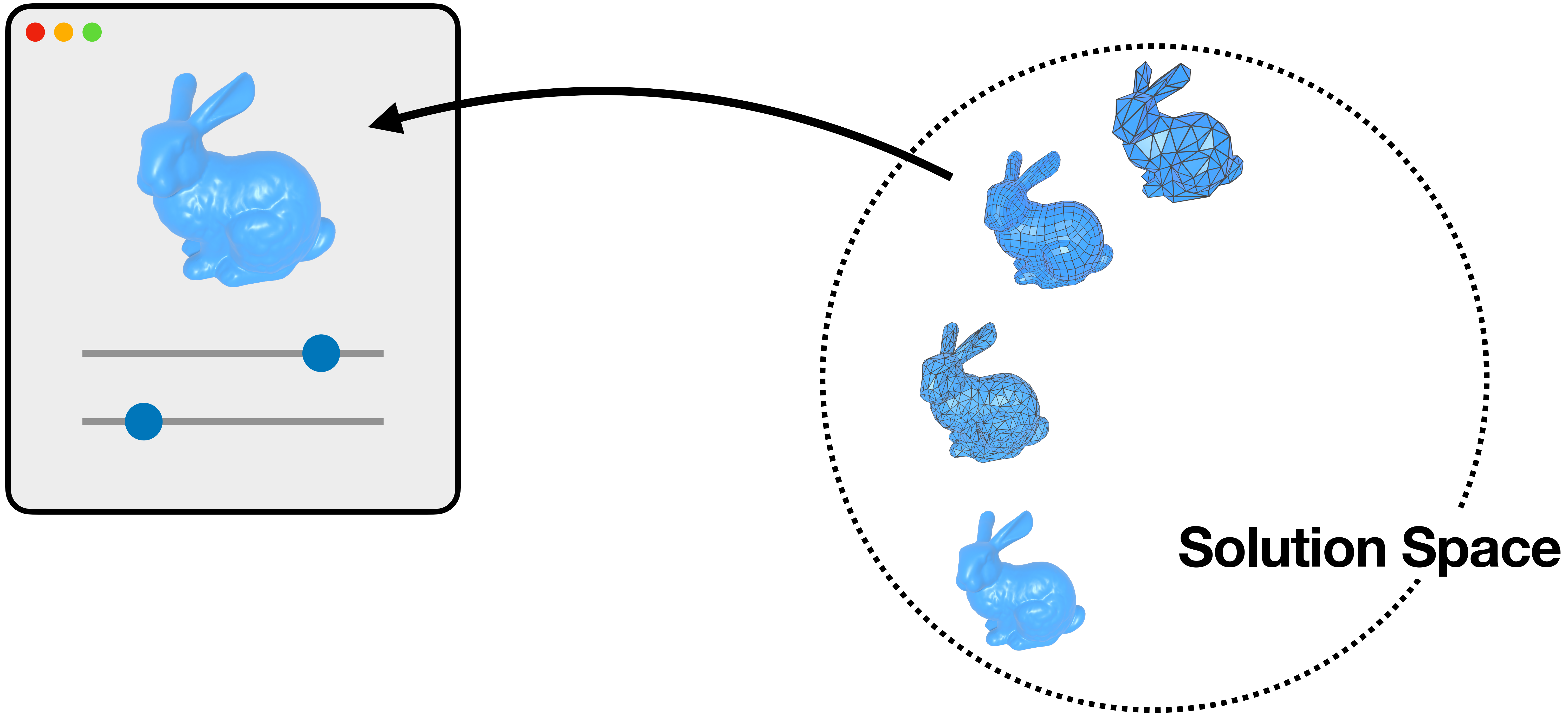

High Rendering Cost
Low Rendering Cost
Perfect Shape Quality
Poor Shape Quality

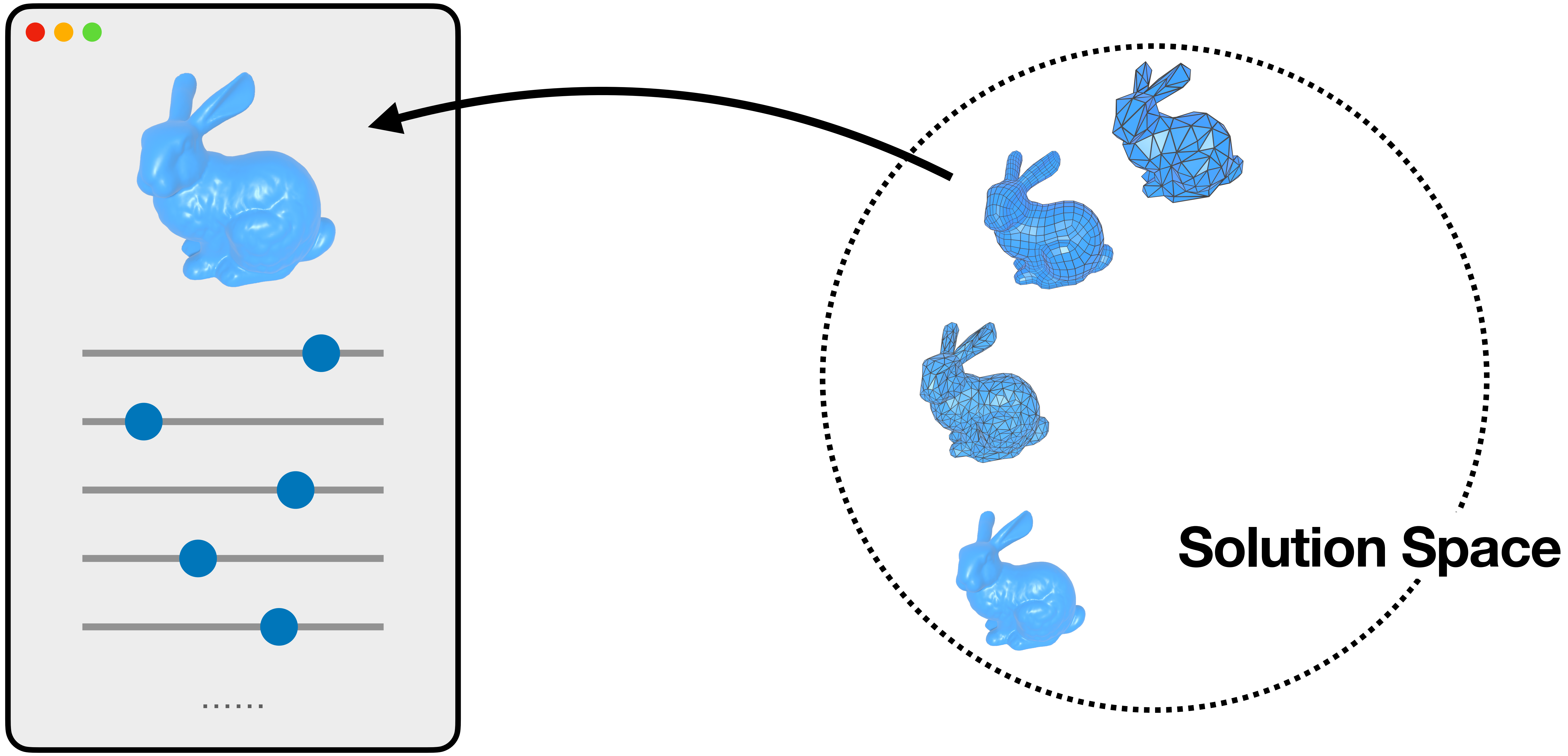

Poor
Shape
Quality
High Rendering Cost
Perfect
Shape
Quality


Poor
Shape
Quality
High Rendering Cost
Perfect
Shape
Quality


Solution Space

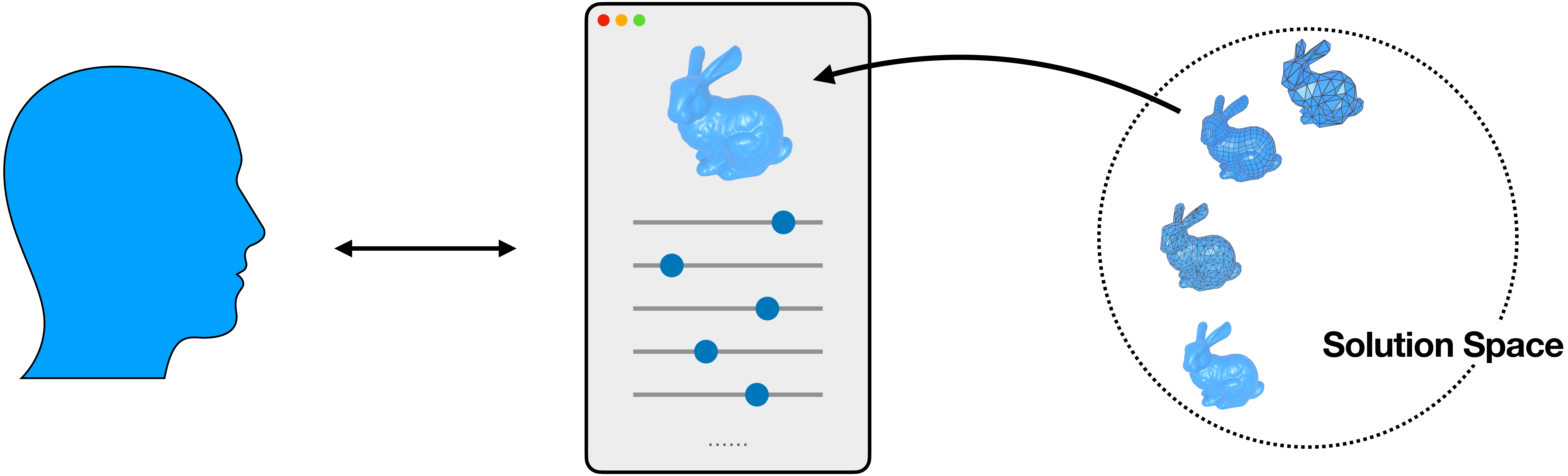

Solution Space


Solution Space



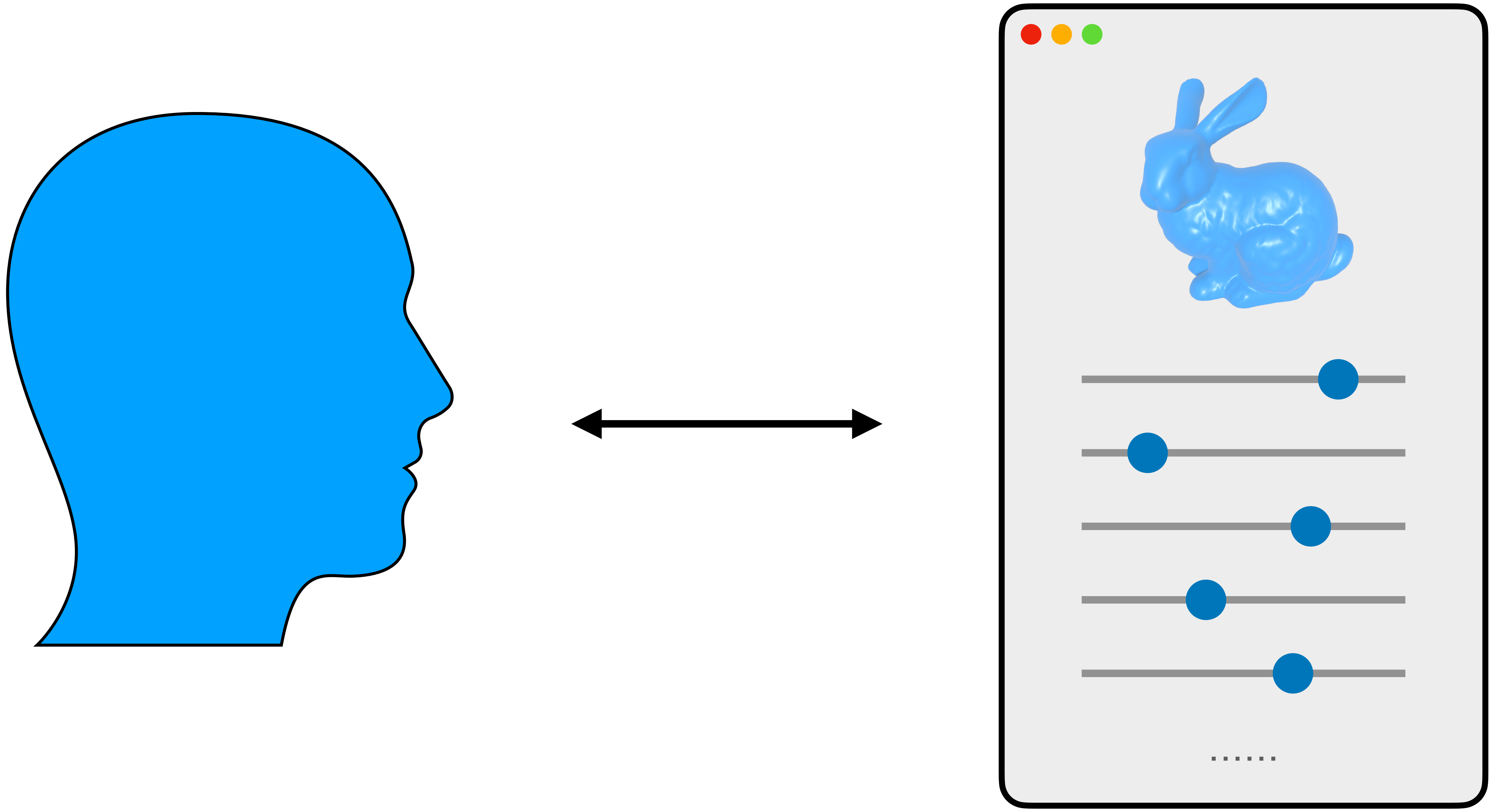
......
Solution Space

Human As the System Operator



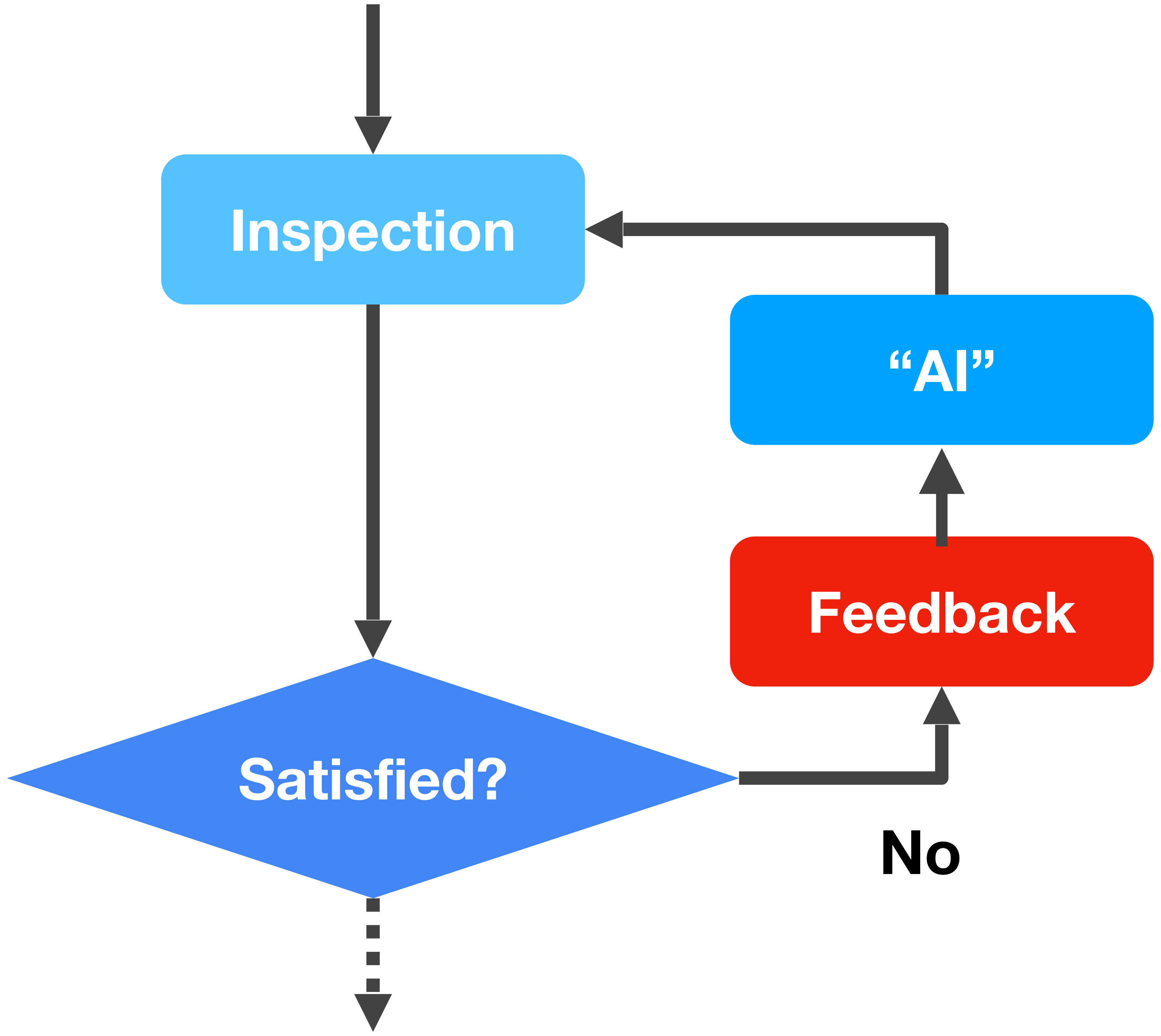
Human-in-the-Loop (HITL) Systems

- 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



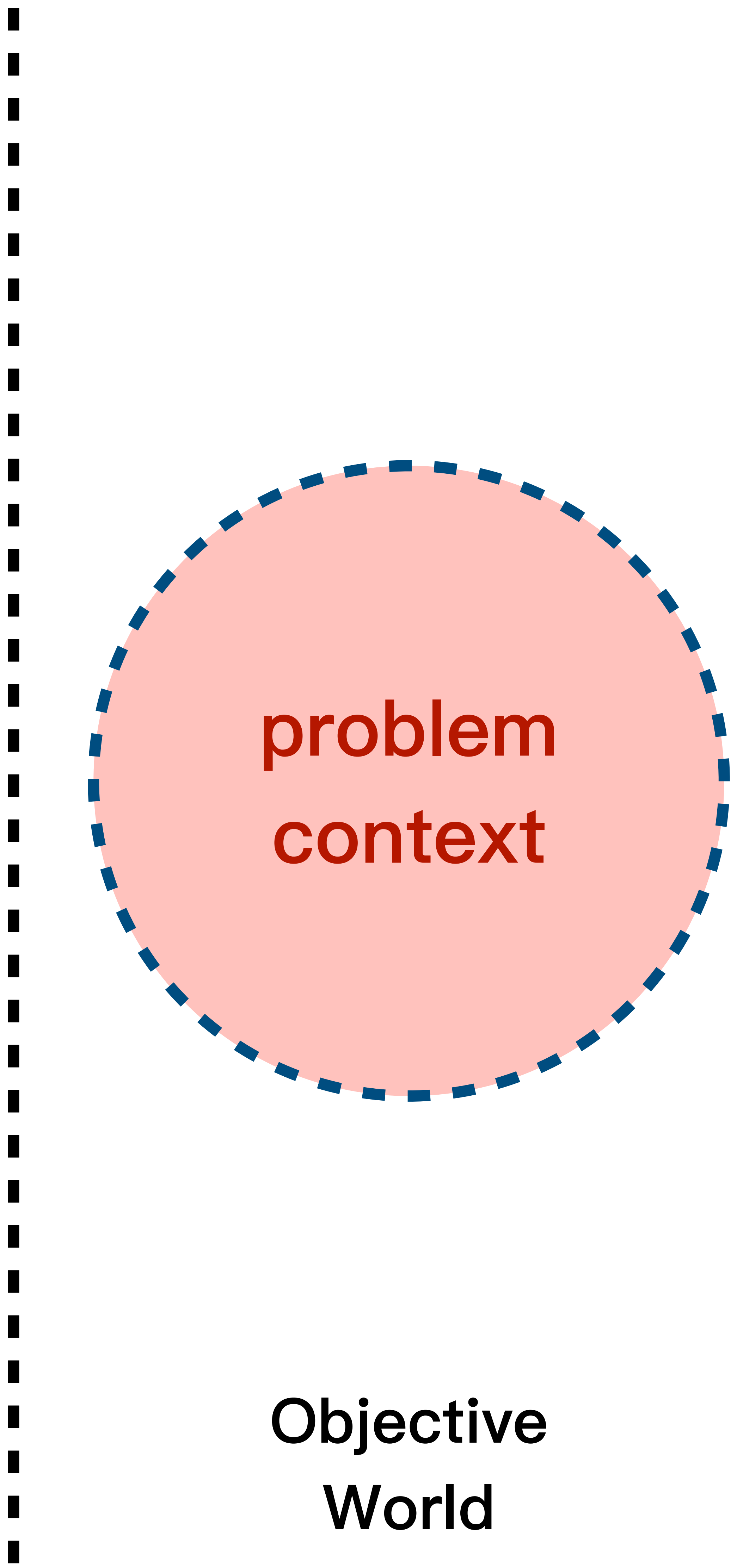
Human-in-the-Loop *Optimization* Systems

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

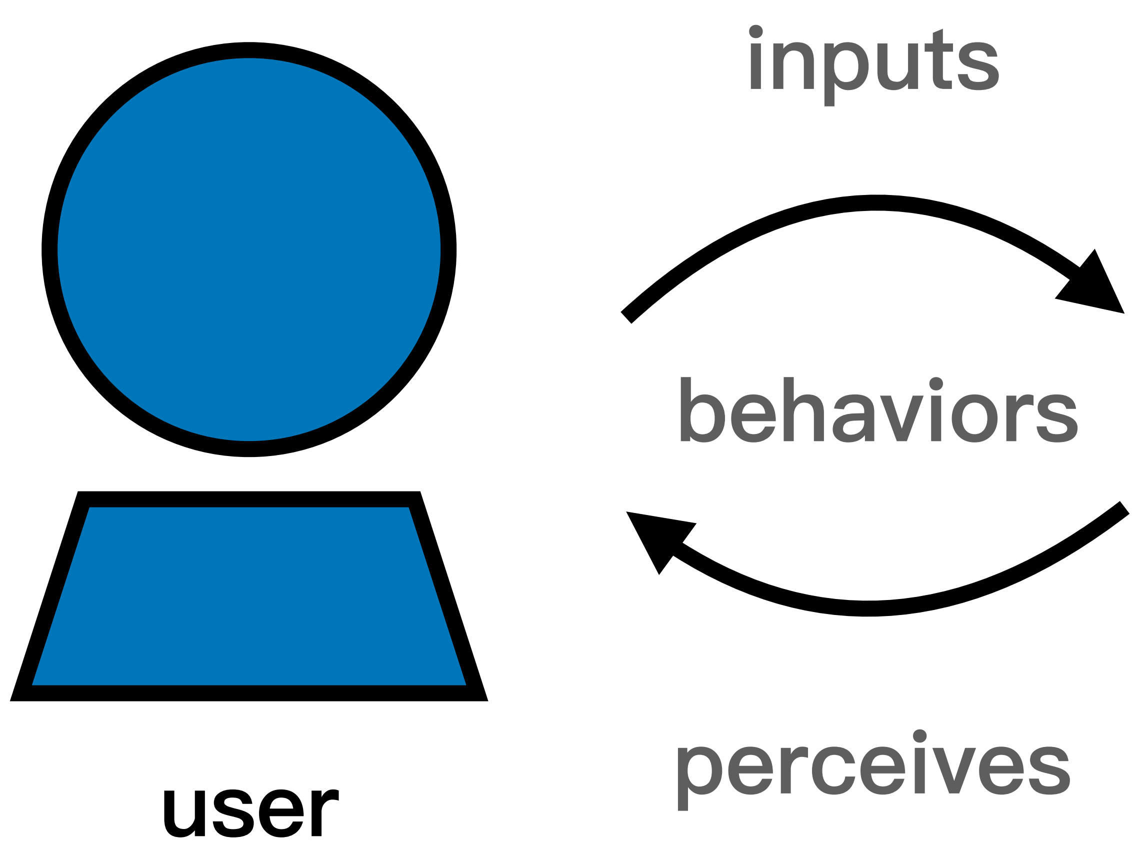
HITL Optimization System Building Blocks



HITL Optimization System Building Blocks

HITL Optimization System Building Blocks

HITL Optimization System Building Blocks



HITL Optimization System

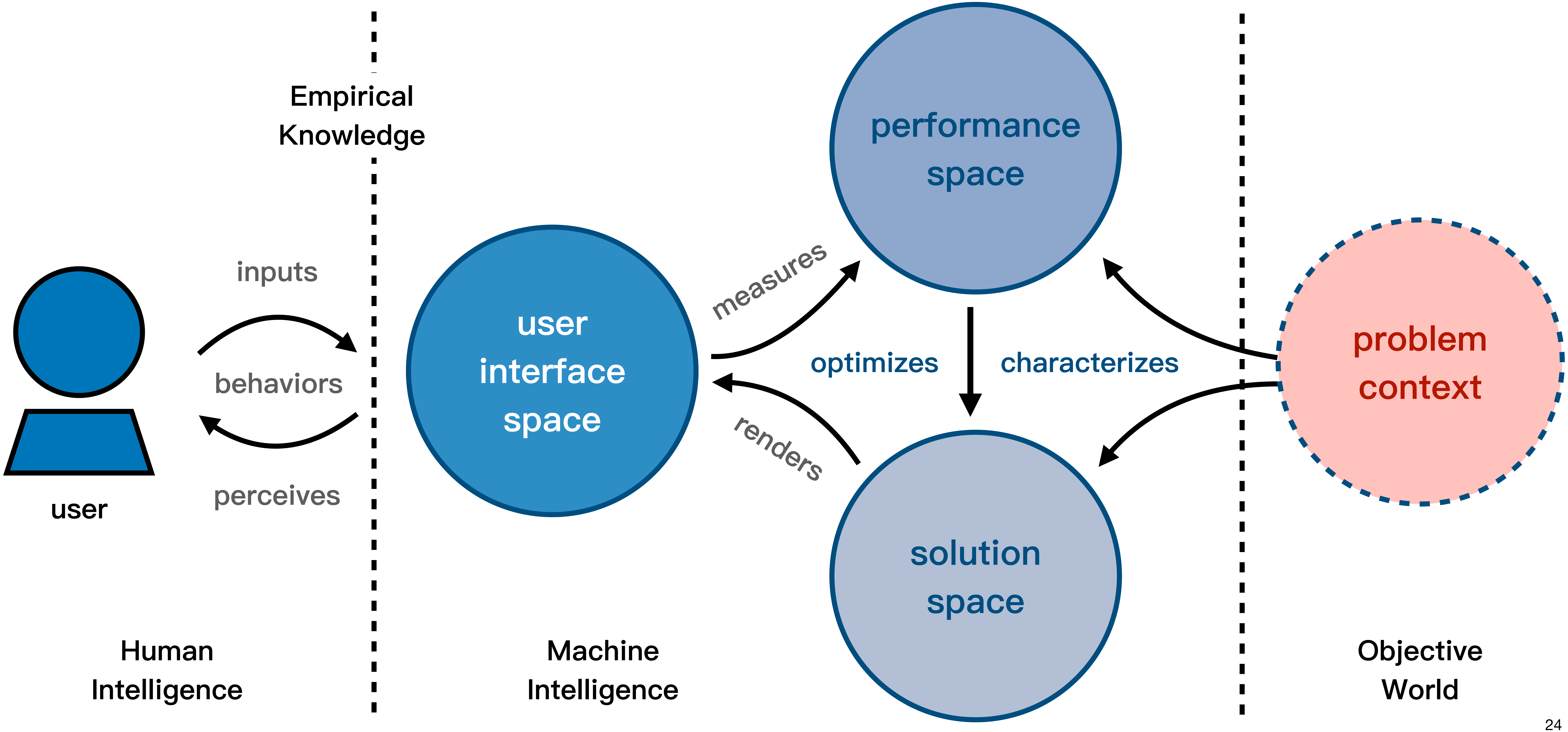
problem context

Human Intelligence

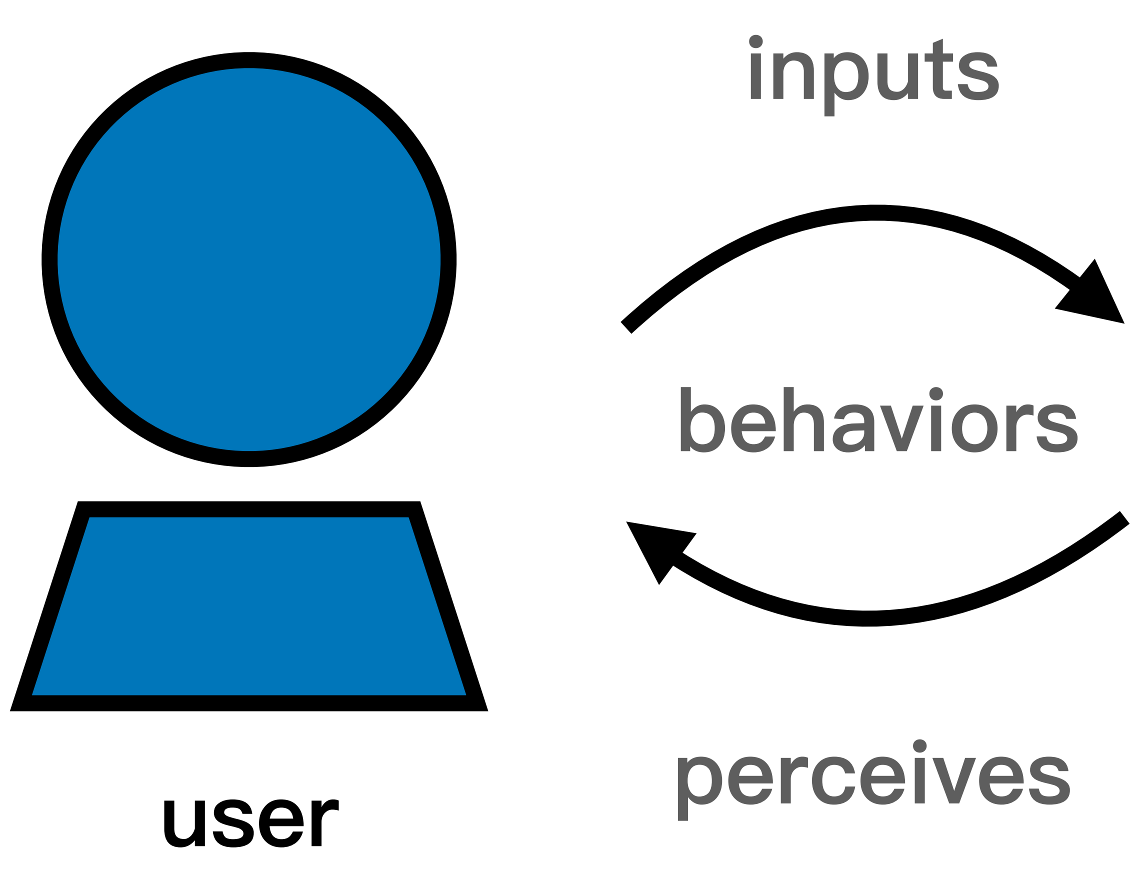
Objective World

HITL Optimization System Building Blocks

HITL Optimization System Building Blocks



HITL Optimization System Building Blocks



Human
Intelligence

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

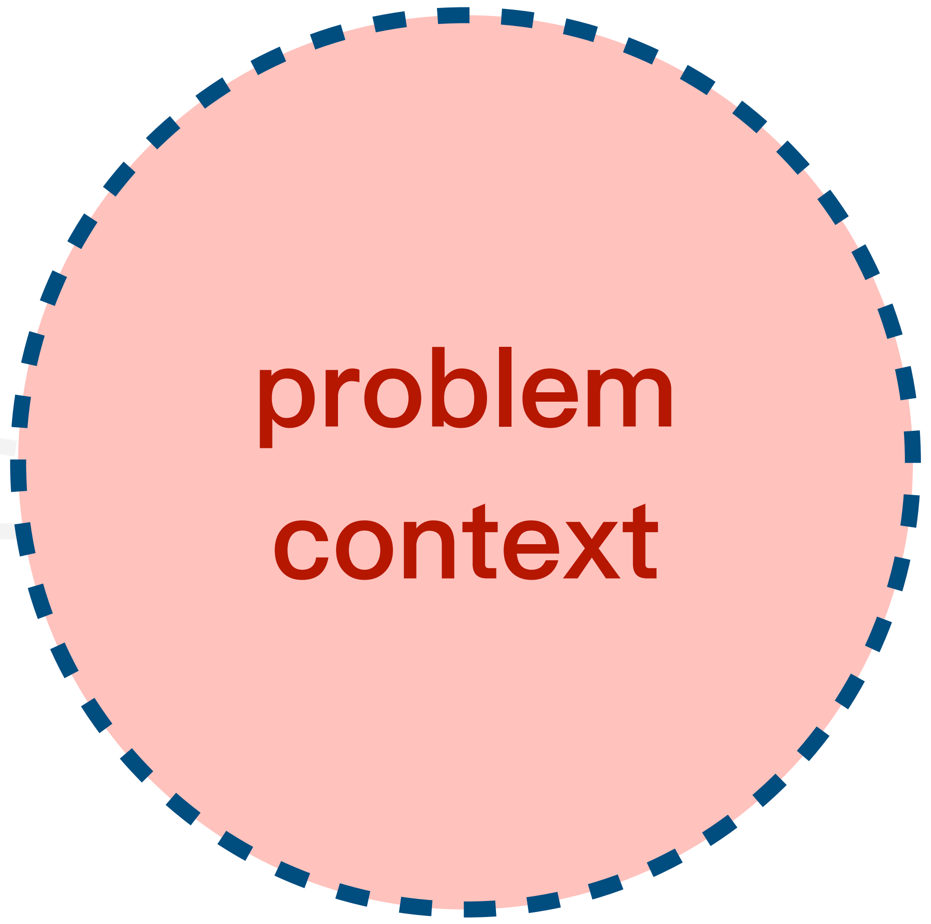
RQ2 Performance metrics What are the relevant performance metrics, and how can we measure them to facilitate the performance comparison between human-in-the-loop systems?

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?

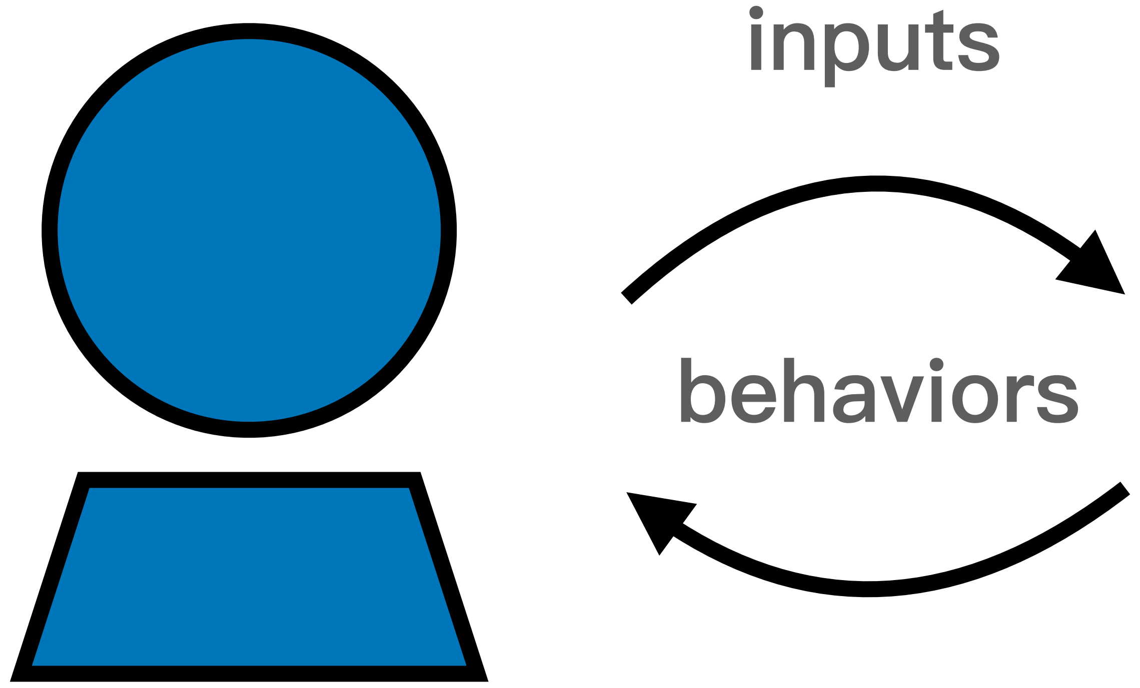
RQ6 Objective alignment How can we identify the alignment of objectives between the human user and the machine system?



Objective
World

Empirical Study I

The Human in the Infinite Loop [Ou et al. MuC’ 22]



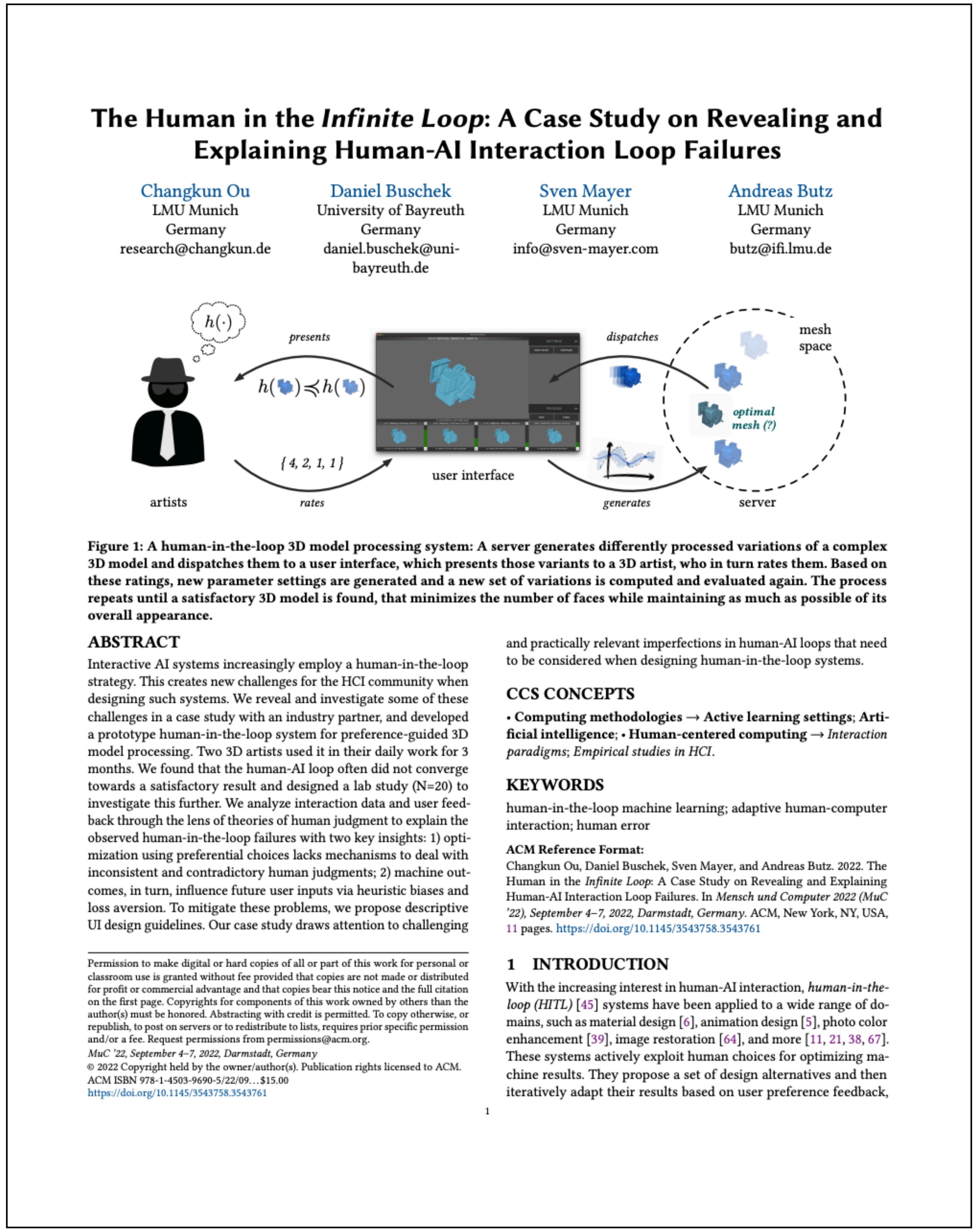
inputs

behaviors

perceives

user

Human
Intelligence



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

Changkun Ou, LMU Munich, Germany, research@changkun.de

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

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

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

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

ABSTRACT

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

CCS CONCEPTS

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

KEYWORDS

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

ACM Reference Format:

Changkun Ou, Daniel Buschek, Sven Mayer, and Andreas Butz. 2022. The Human in the *Infinite Loop*: A Case Study on Revealing and Explaining Human-AI Interaction Loop Failures. In *Mensch und Computer 2022 (MuC ’22), September 4–7, 2022, Darmstadt, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543758.3543761

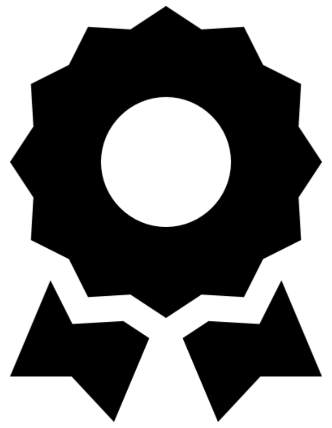
1 INTRODUCTION

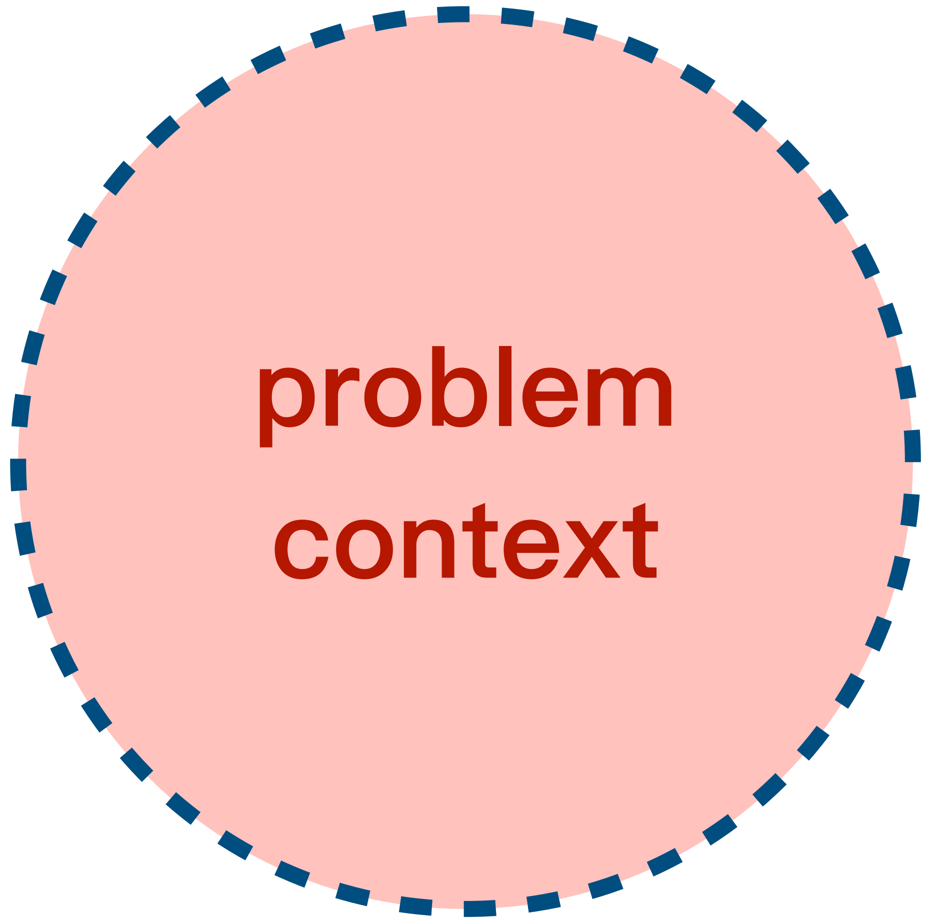
With the increasing interest in human-AI interaction, *human-in-the-loop (HITL)* [45] systems have been applied to a wide range of domains, such as material design [6], animation design [5], photo color enhancement [39], image restoration [64], and more [11, 21, 38, 67]. These systems actively exploit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

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

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

1





Objective
World

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]

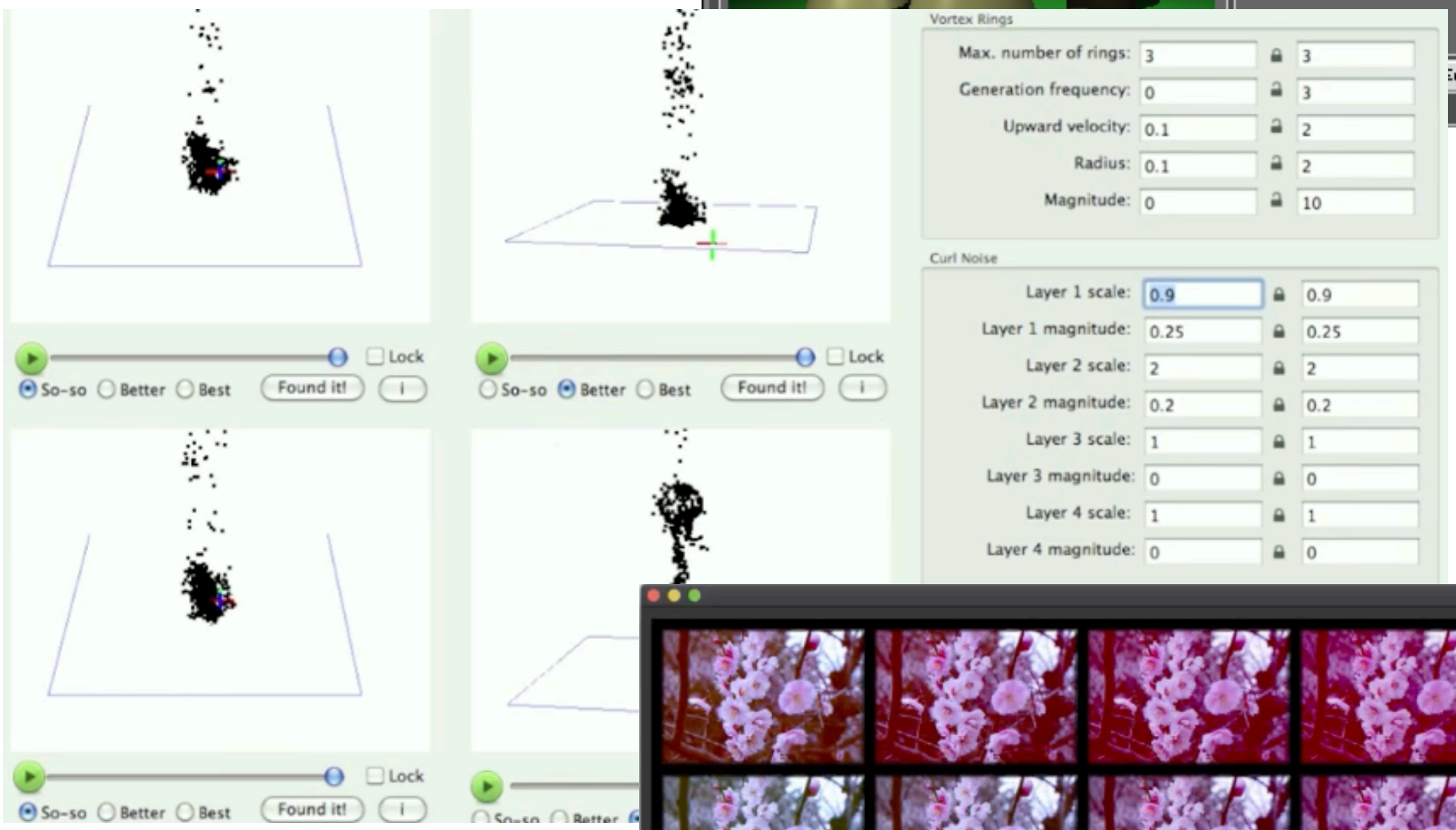
Material design [Brochu et al. 2007]

Animation [Brochu et al. 2010]

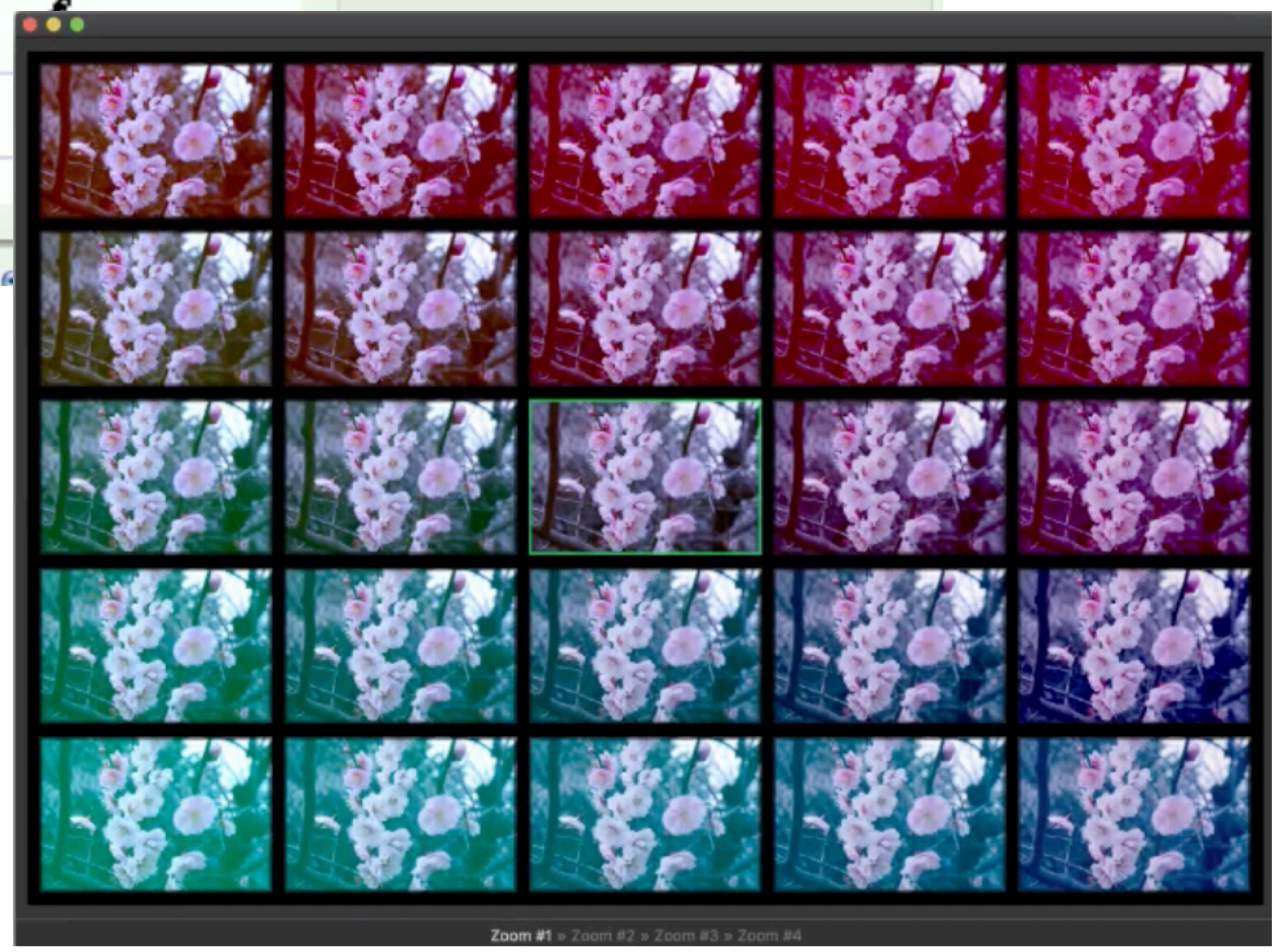
Illumination [Marks et al. 1997]

…

[Marks et al. 1997]

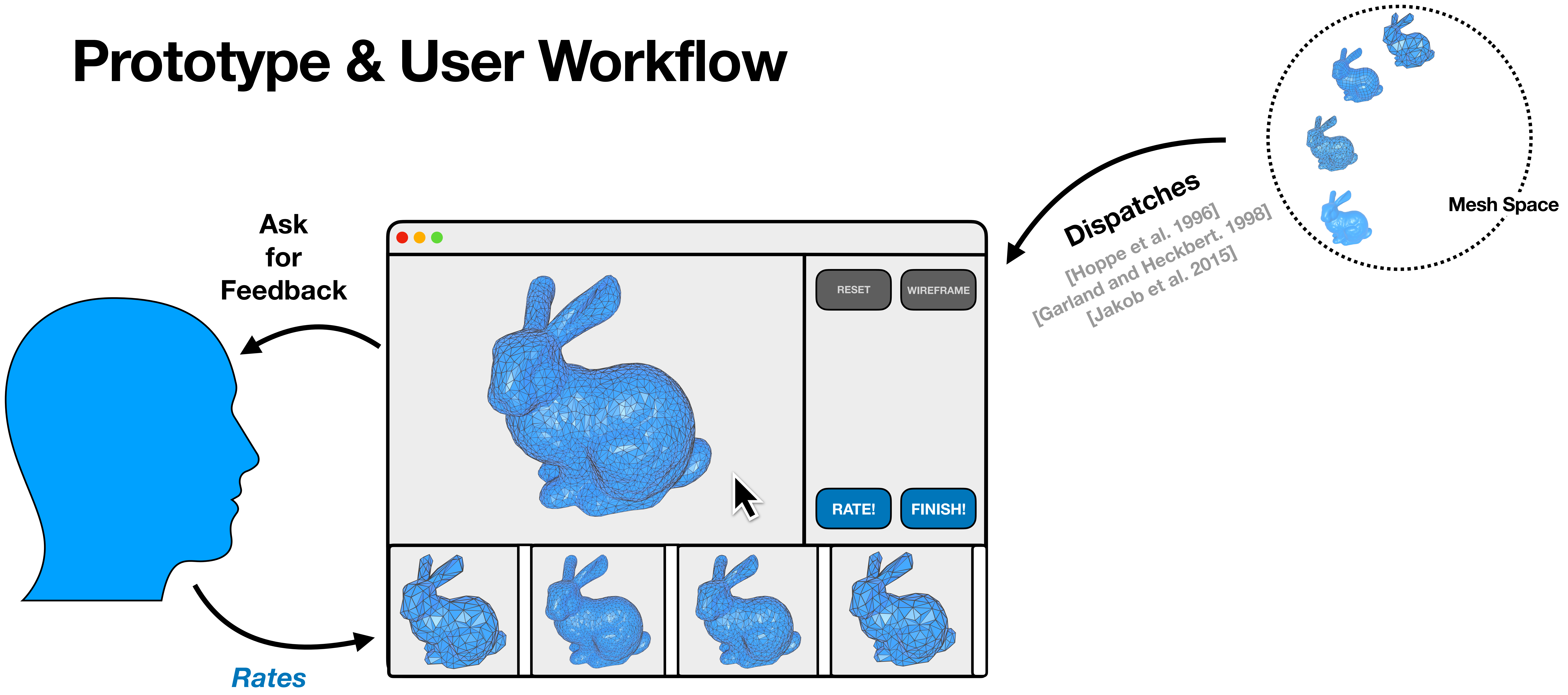


[Brochu et al. 2007]

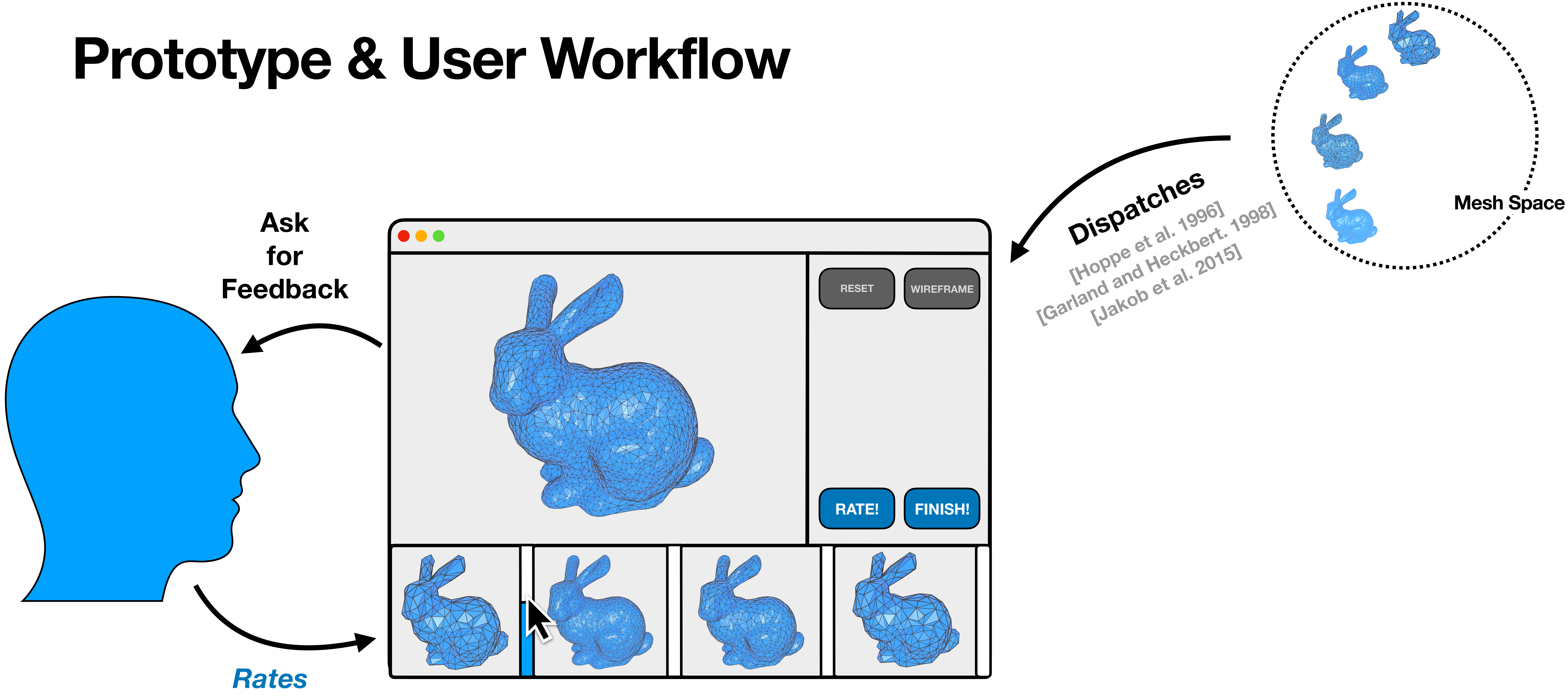


[Koyama et al. 2020]

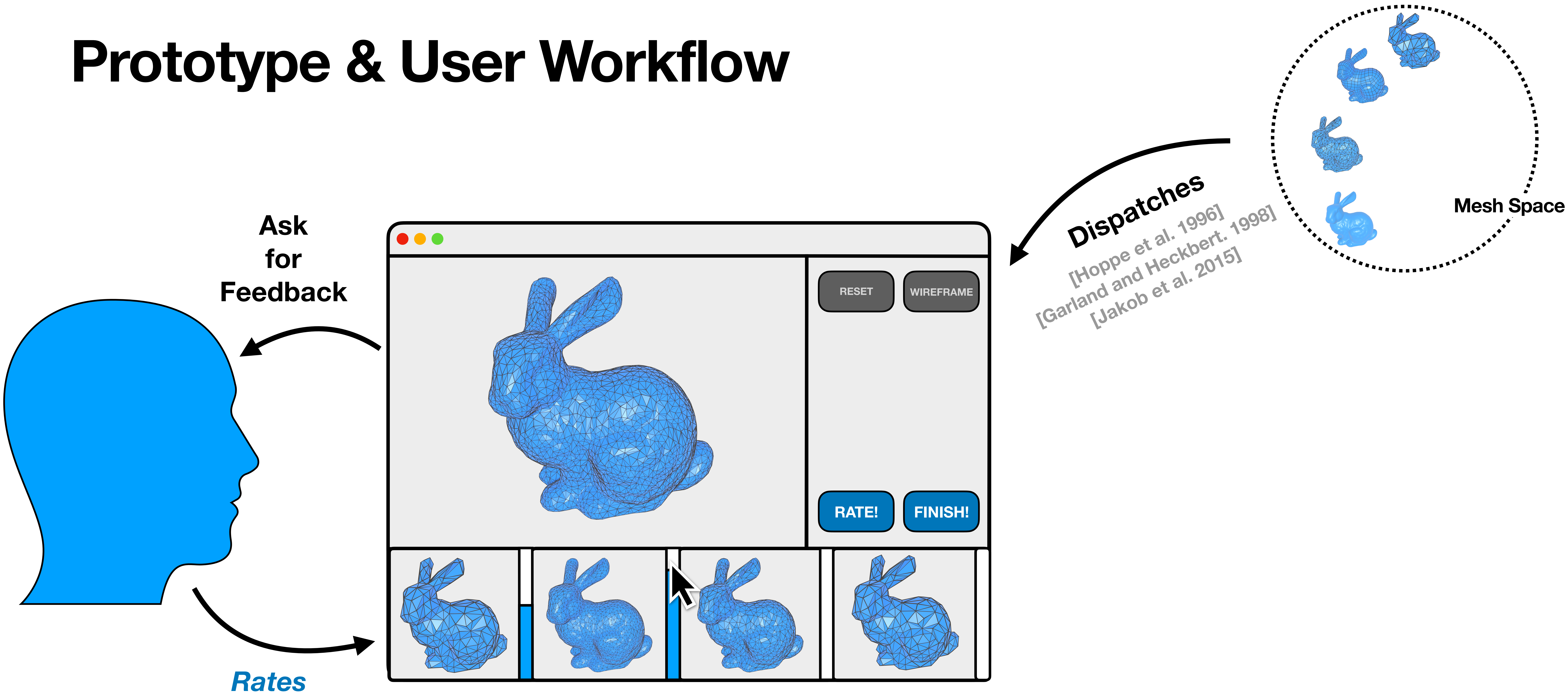
Prototype & User Workflow



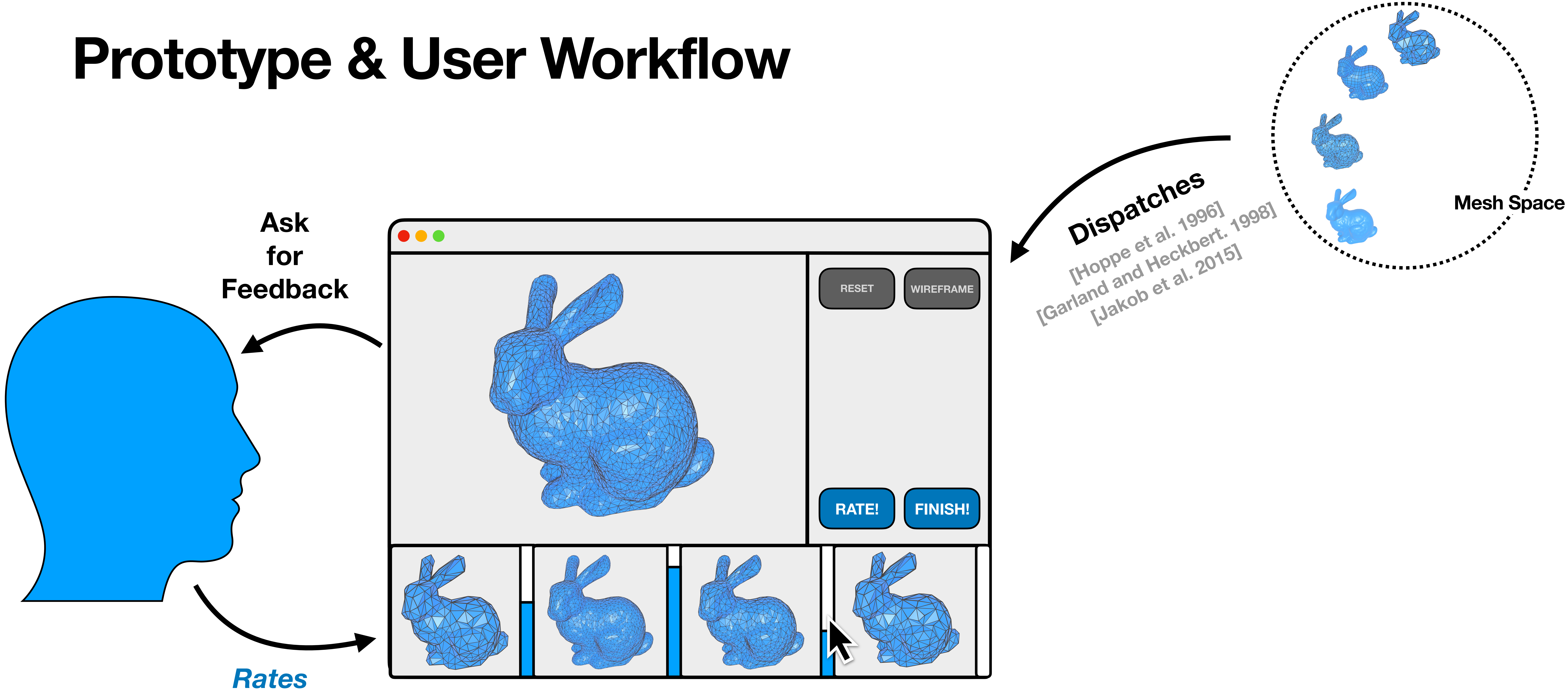
Prototype & User Workflow



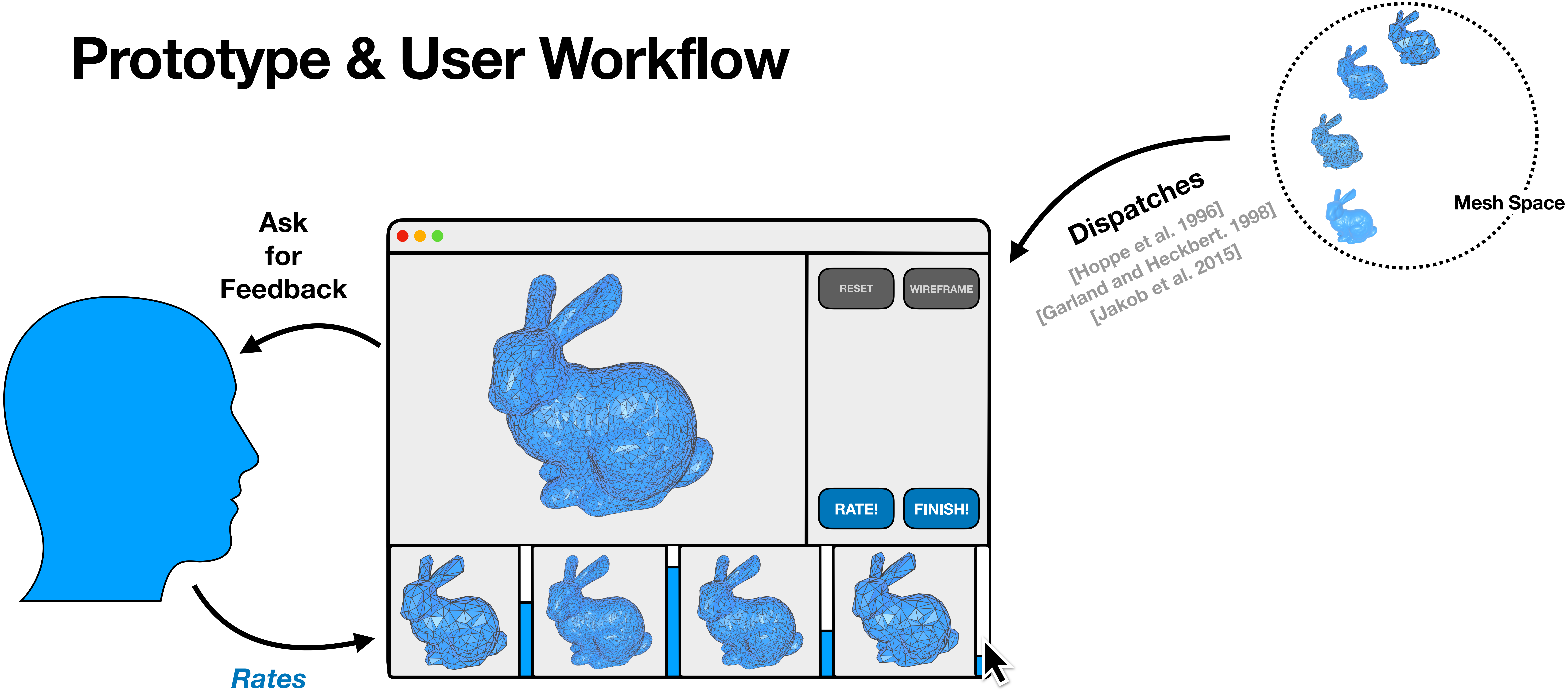
Prototype & User Workflow



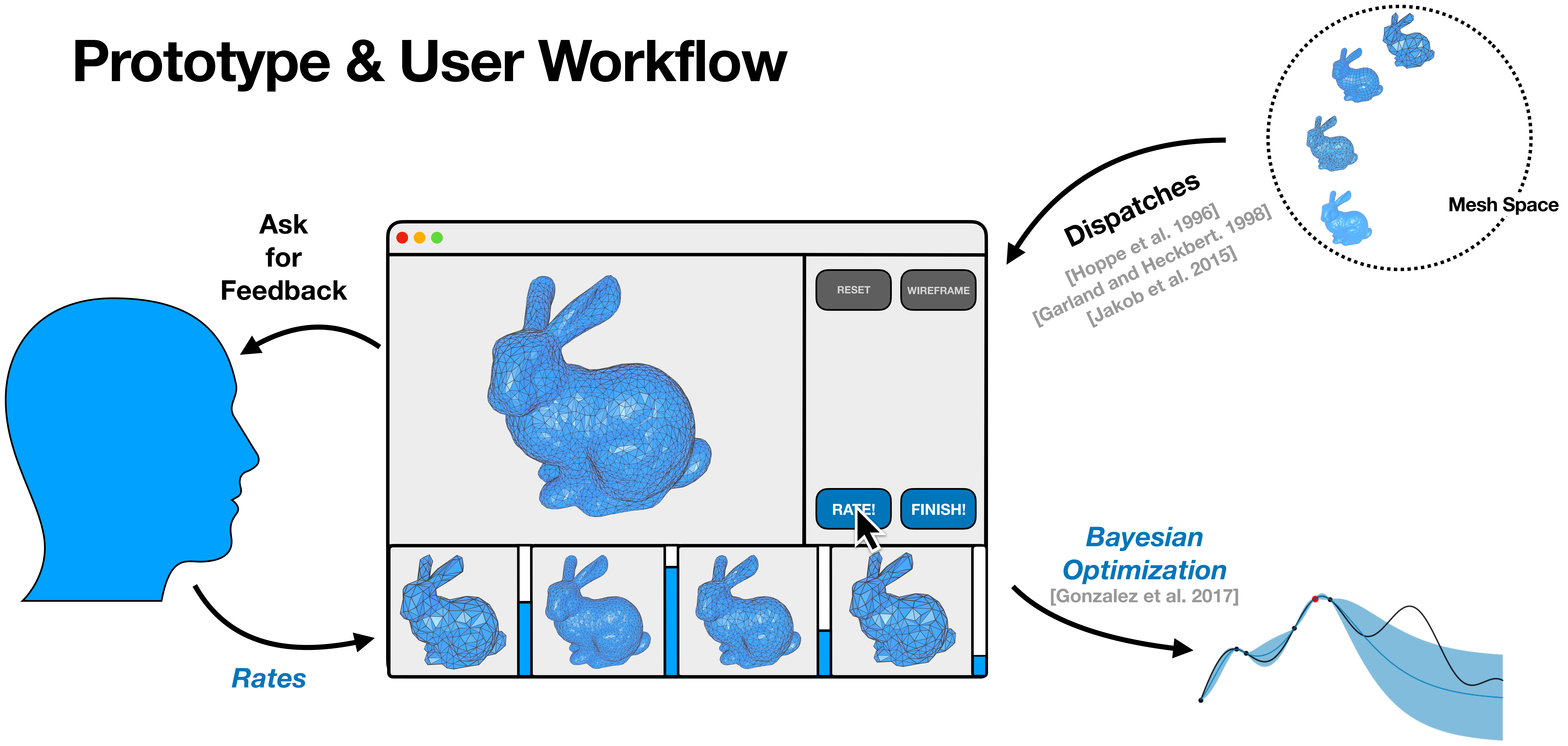
Prototype & User Workflow



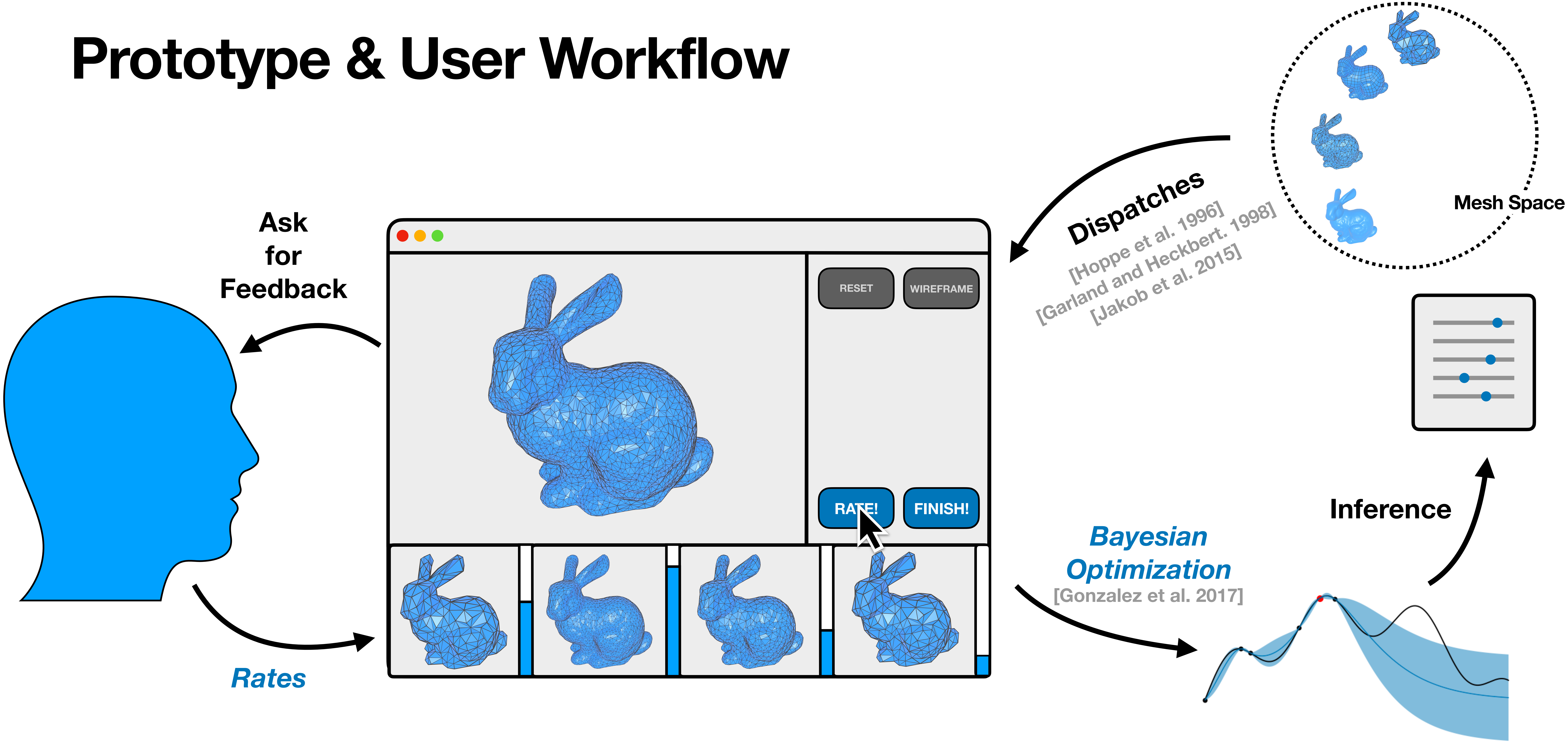
Prototype & User Workflow



Prototype & User Workflow



Prototype & User Workflow



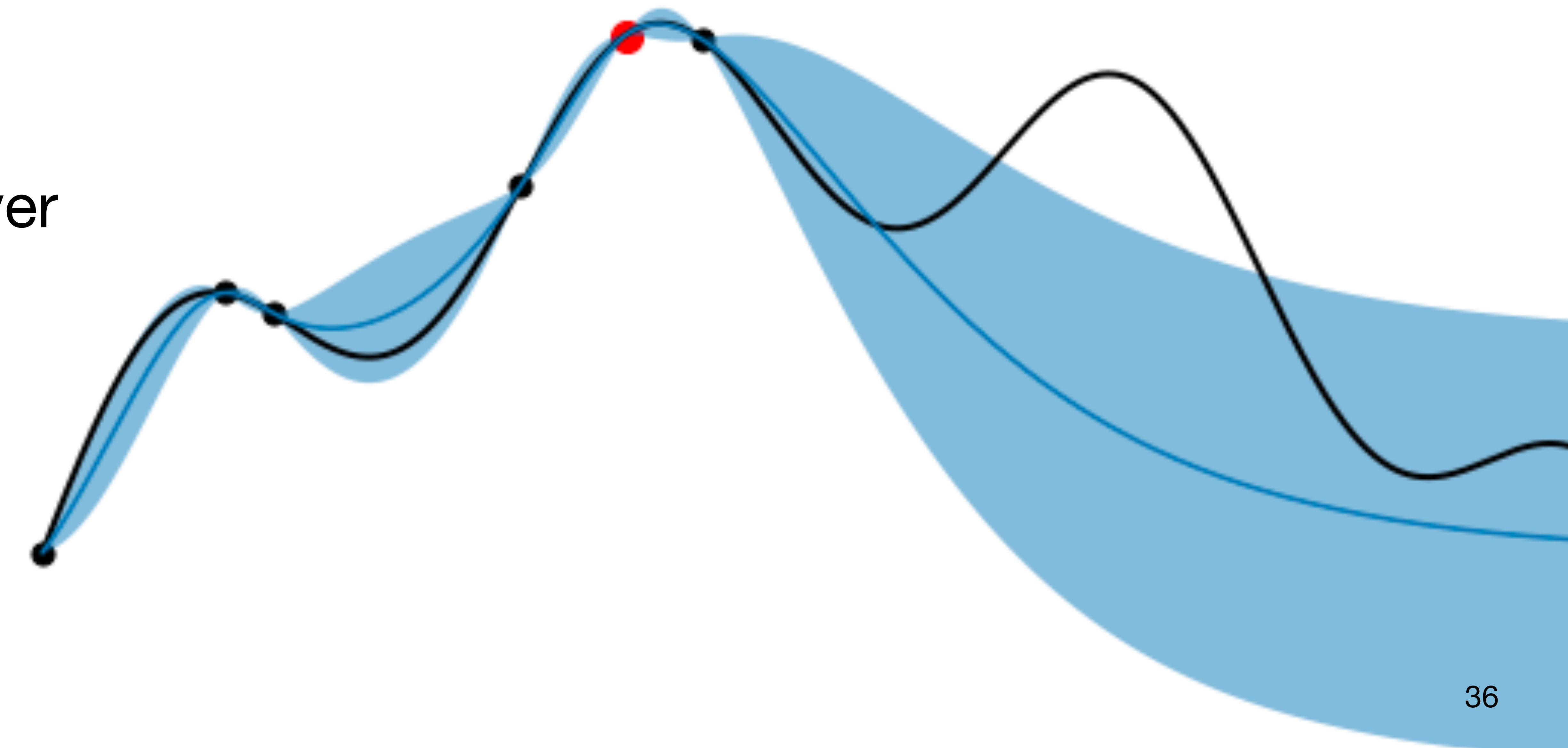
Bayesian Optimization (BO)

Bayesian optimization [Mockus, 1978] aims to find $\mathbf{x}^* = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}}\, f(\mathbf{x})$

With a initial dataset $\mathcal{D}_0 = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$

1. Choose some **prior measure** over the space of possible objectives $f(\mathbf{x}) \sim GP(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$
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** $\mathbf{x}_{t+1} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}}\, AF(\mathbf{x}; \mathcal{D}_t)$
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 $f([\mathbf{x}, \mathbf{x}']) = g(\mathbf{x}') - g(\mathbf{x})$ that defines a **preference function**

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

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

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

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

$$[\mathbf{x}_{j+1}, \mathbf{x}'_{j+1}] = [\operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} \int_{\mathcal{X}} \pi_{\tilde{f}}([\mathbf{x}, \mathbf{x}']; D_j) \mathrm{d}\mathbf{x}', \operatorname*{argmax}_{\mathbf{x}'_*\in\mathcal{X}} \mathbf{V}[\sigma(f_*)|[\mathbf{x}_*, \mathbf{x}'_*], \mathcal{D}_j, \mathbf{x}_* = \mathbf{x}_{j+1}]]$$

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

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)

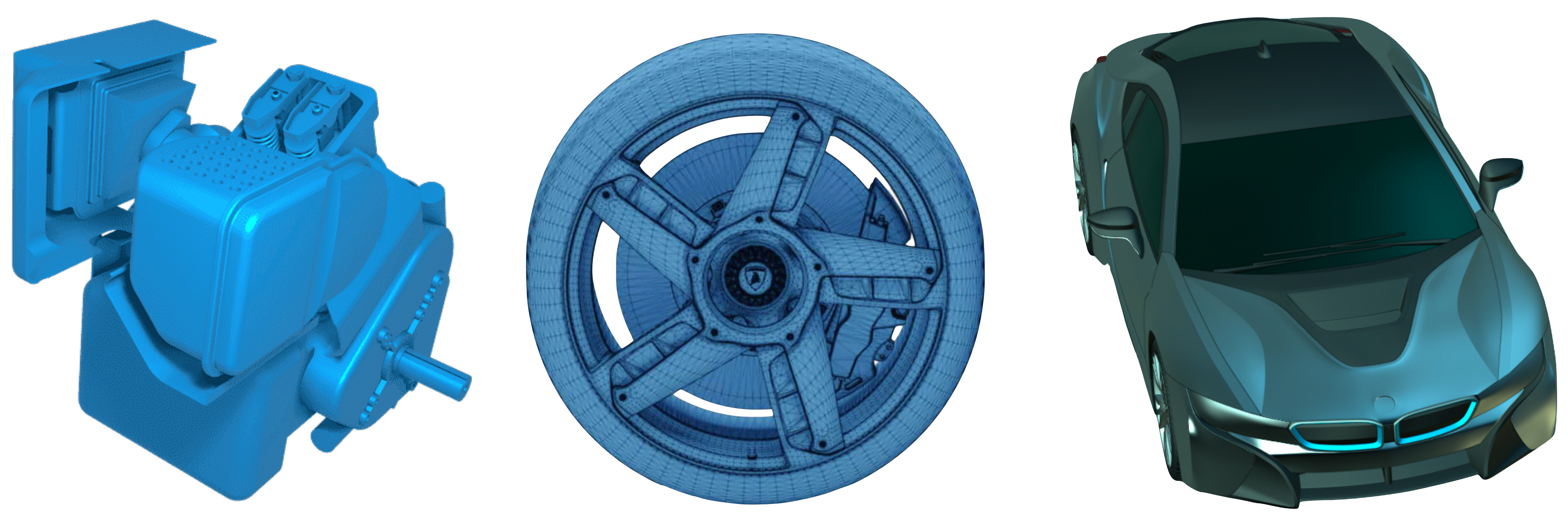
3 months of usage, male technical artists, age range 25-35, both 3 more years experience

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

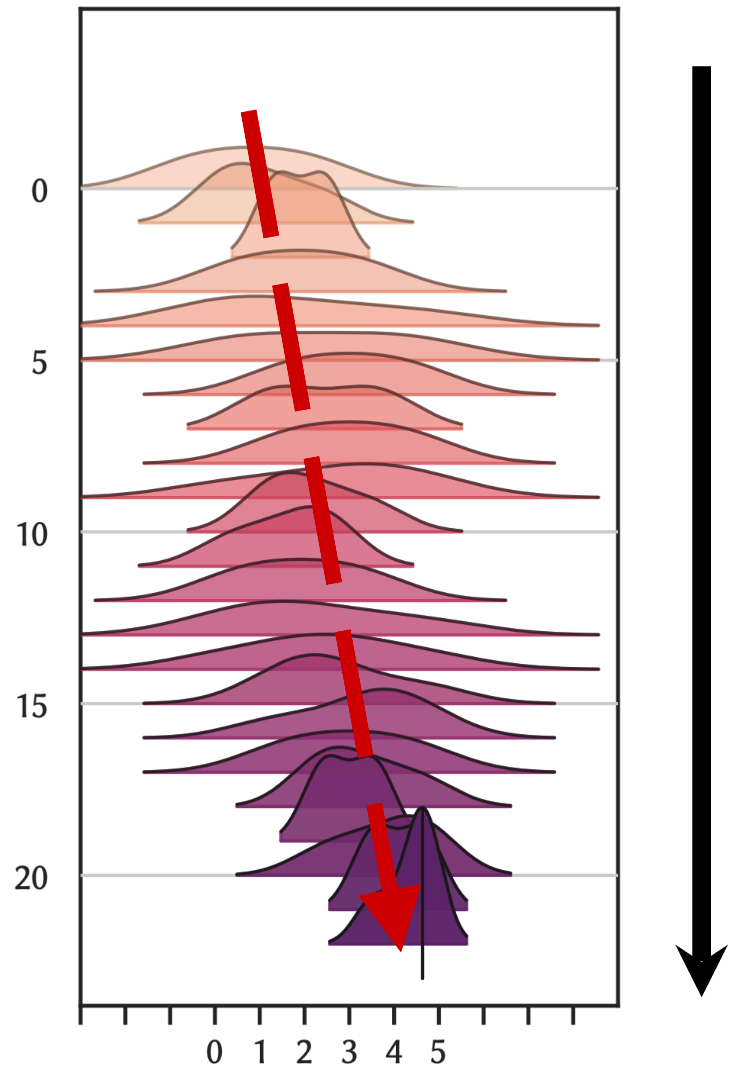


Models Appeared in Field Study

Models Used in Lab 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).

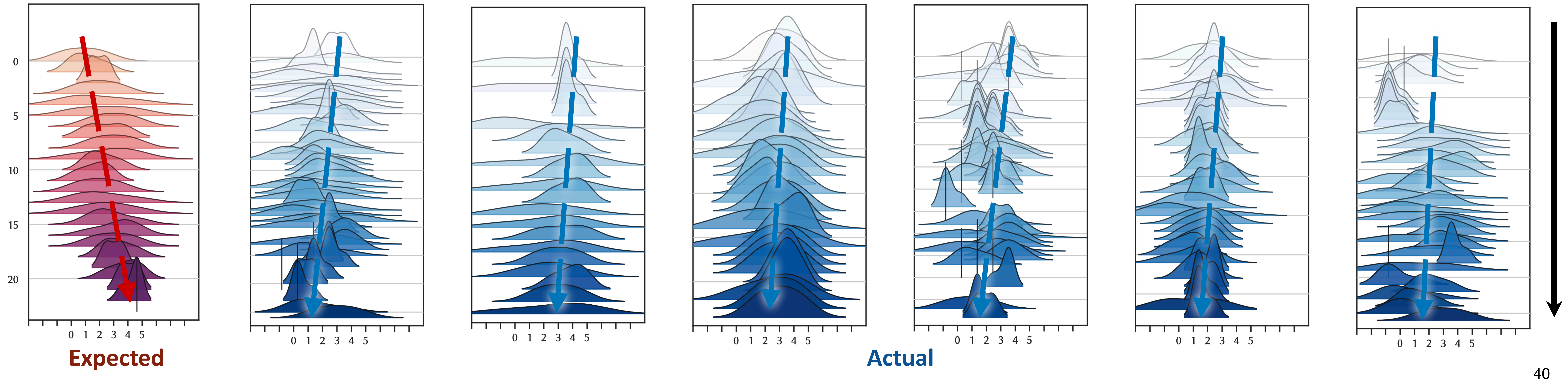


Expected

Observations

Large mismatch between expected and actual ratings

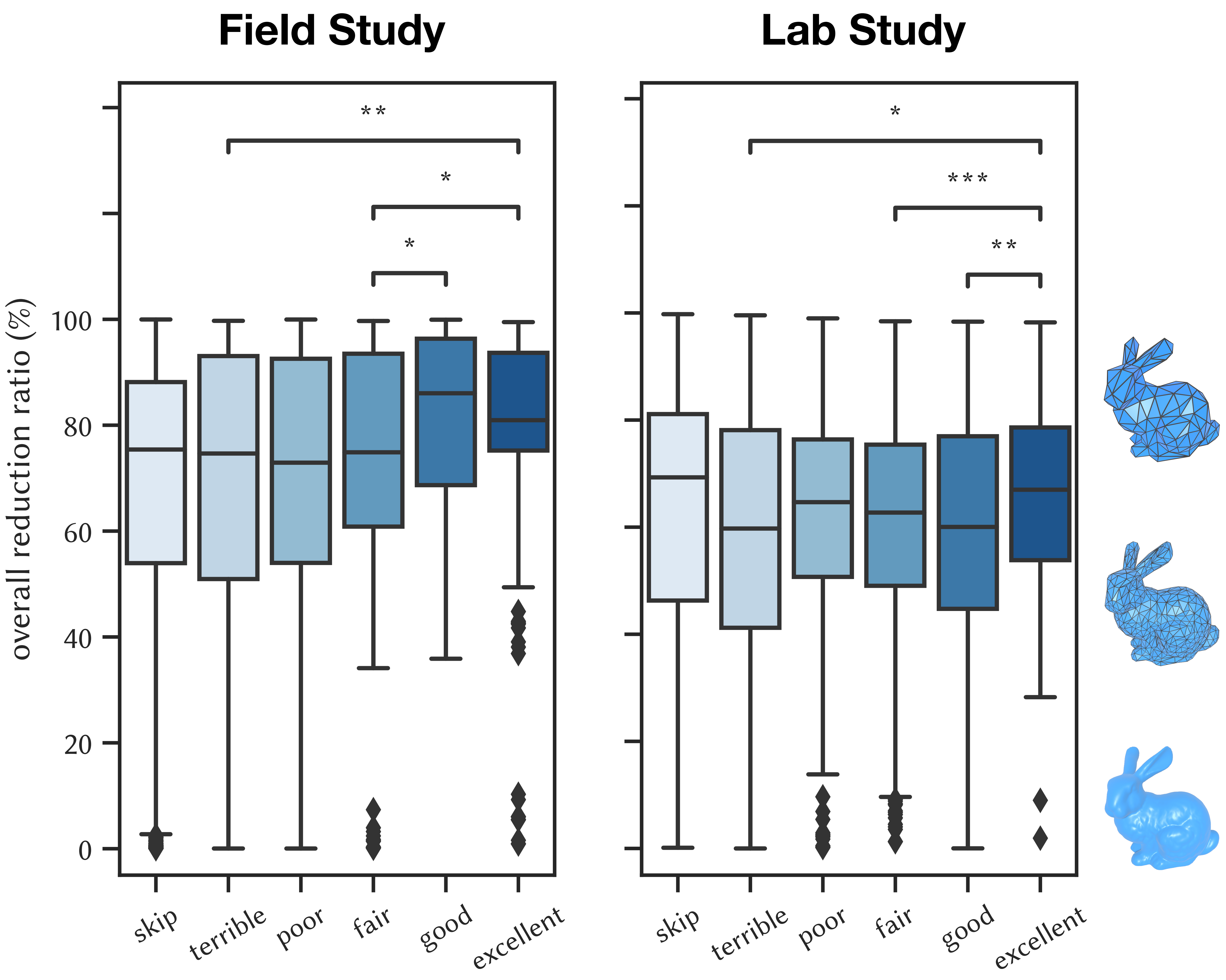
Either non-stationary and decreasing



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

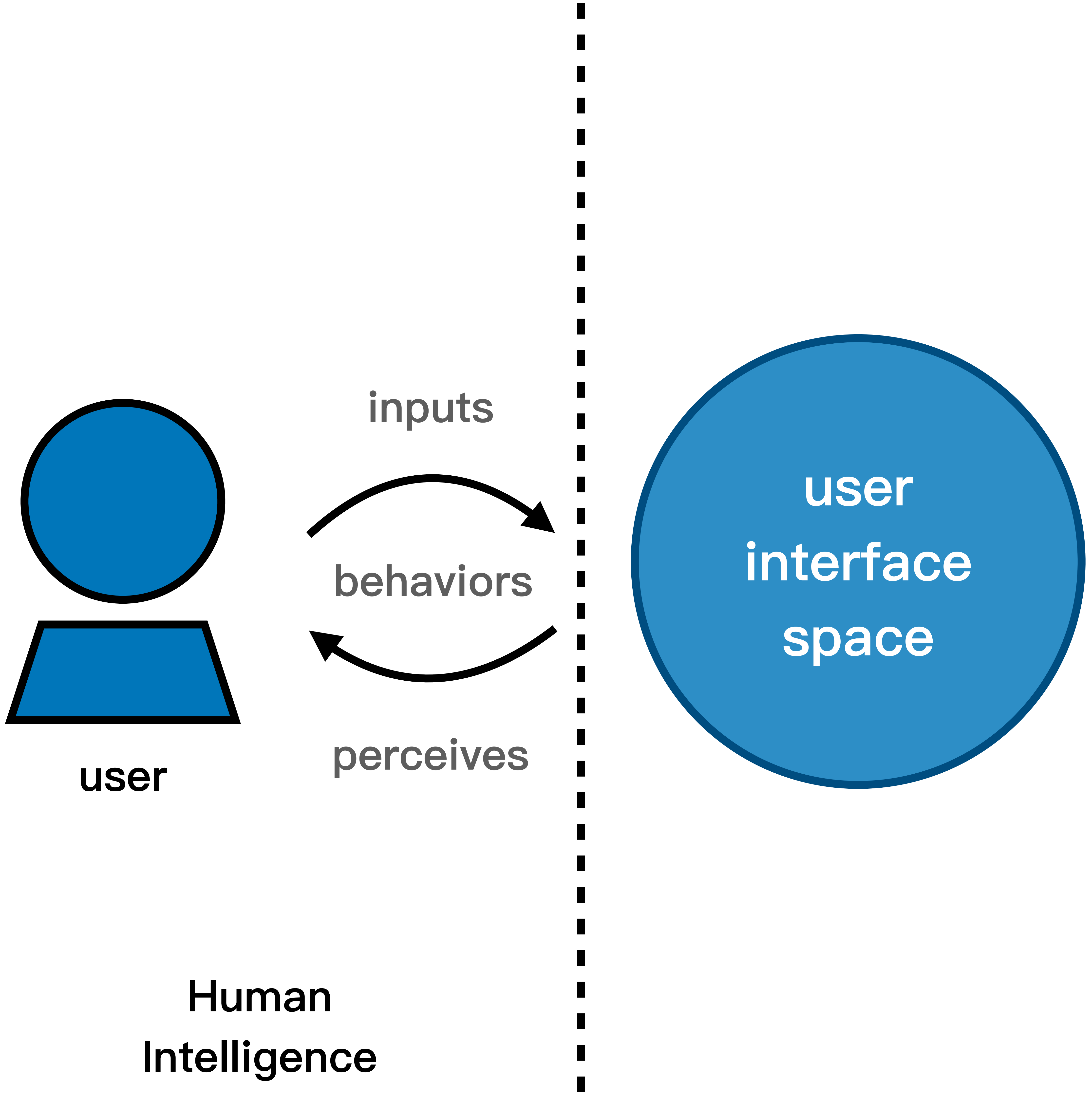
e.g. indicate optimization intention to frame the current context better, mitigating representativeness and availability bias and avoid judging based on previous examples

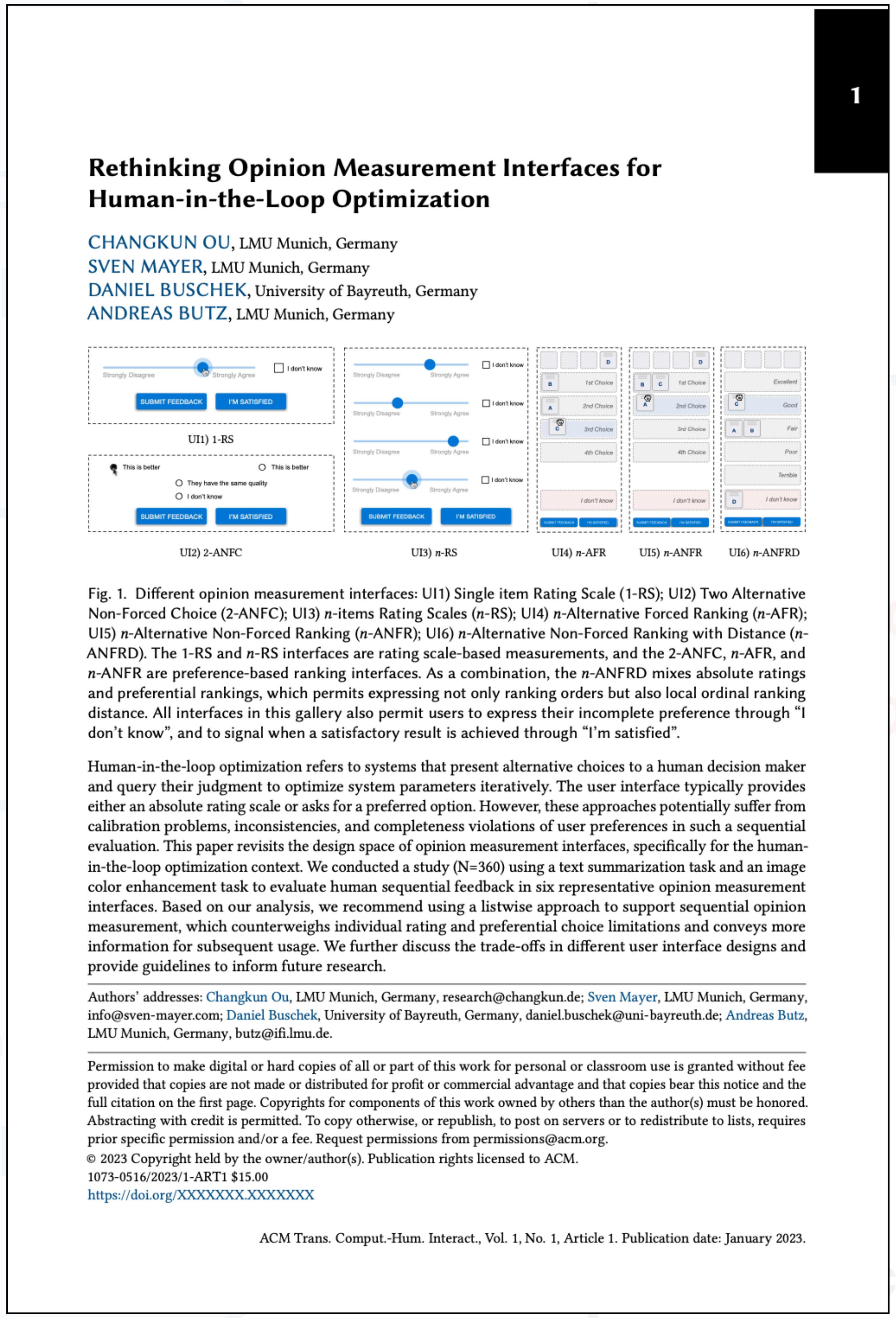
Reduce purely occasional flaws

e.g. present previous results to add consistency check

Empirical Study II

Rethinking Opinion Measurement Interfaces [Ou et al. ToCHI' 2?]





Rethinking Opinion Measurement Interfaces for Human-in-the-Loop Optimization

CHANGKUN OU, LMU Munich, Germany
SVEN MAYER, LMU Munich, Germany
DANIEL BUSCHEK, University of Bayreuth, Germany
ANDREAS BUTZ, LMU Munich, Germany

UI1) 1-RS UI2) 2-ANFC UI3) n-RS UI4) n-AFR UI5) n-ANFR UI6) n-ANFRD

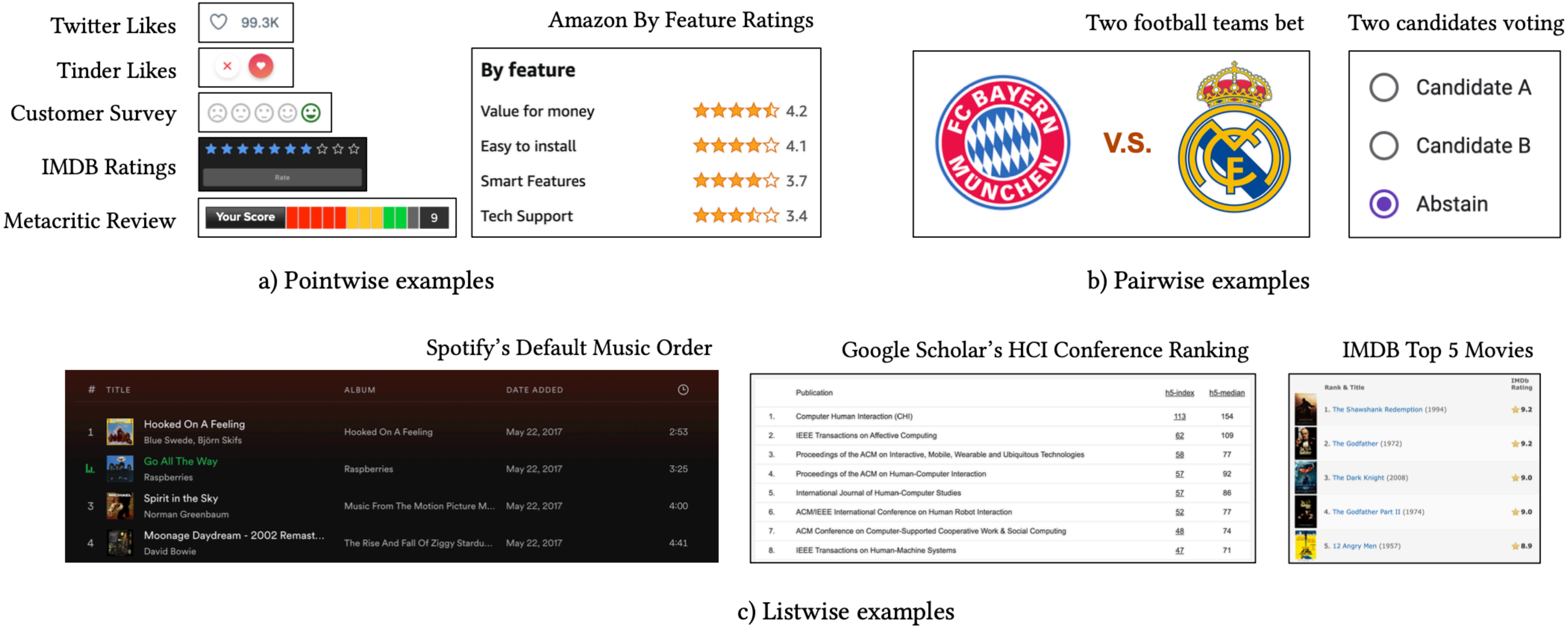
Fig. 1. Different opinion measurement interfaces: UI1) Single item Rating Scale (1-RS); UI2) Two Alternative Non-Forced Choice (2-ANFC); UI3) n-items Rating Scales (n-RS); UI4) n-Alternative Forced Ranking (n-AFR); UI5) n-Alternative Non-Forced Ranking (n-ANFR); UI6) n-Alternative Non-Forced Ranking 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".

Human-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 asks for a preferred option. However, these approaches potentially suffer from calibration problems, inconsistencies, and completeness violations of user preferences in such a sequential evaluation. This paper revisits the design space of opinion measurement interfaces, specifically for the human-in-the-loop optimization context. We conducted a study (N=360) using a text summarization task and an image color enhancement task to evaluate human sequential feedback in six representative opinion measurement interfaces. Based on our analysis, we recommend using a listwise approach to support sequential opinion measurement, which counterweighs individual rating and preferential choice limitations and conveys more information for subsequent usage. We further discuss the trade-offs in different user interface designs and provide guidelines to inform future research.

Authors' addresses: Changkun Ou, LMU Munich, Germany, research@changkun.de; Sven Mayer, LMU Munich, Germany, info@sven-mayer.com; Daniel Buschek, University of Bayreuth, Germany, daniel.buschek@uni-bayreuth.de; Andreas Butz, LMU Munich, Germany, butz@ifi.lmu.de.

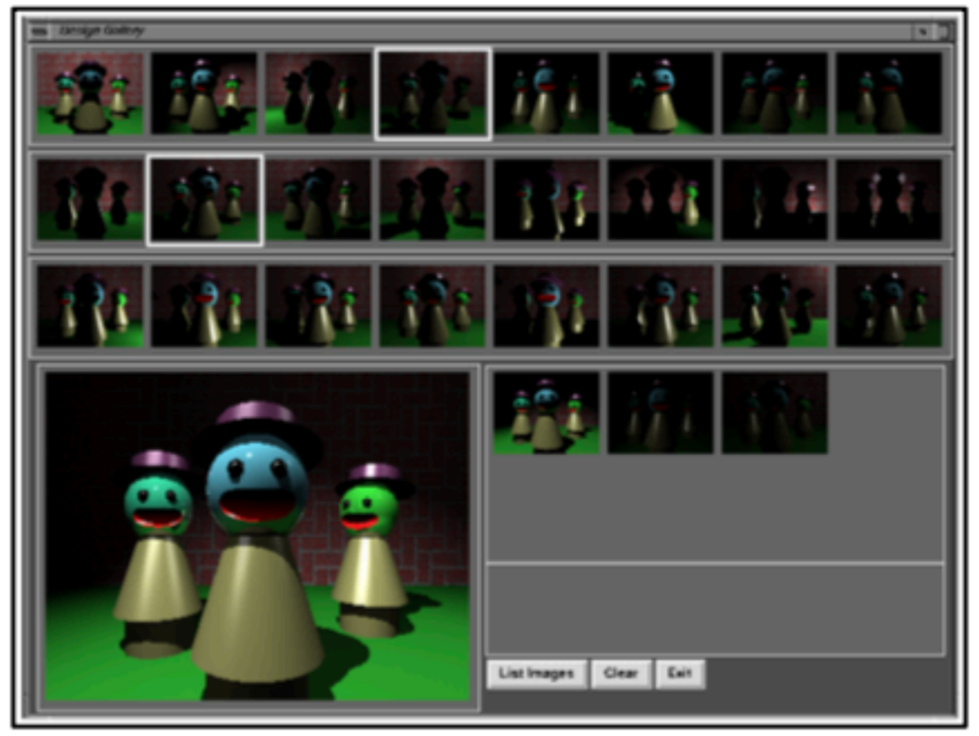
1073-0516/2023/1-ART1 $15.00
https://doi.org/XXXXXXX.XXXXXXX

Opinion Measurement Interfaces

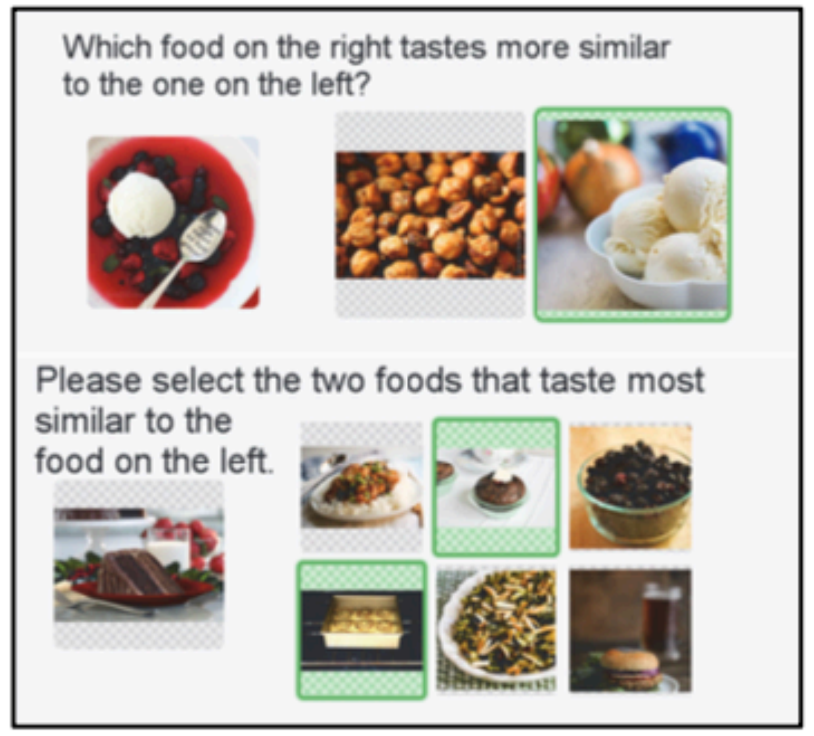


Opinion Measurement Interfaces for HITL Systems

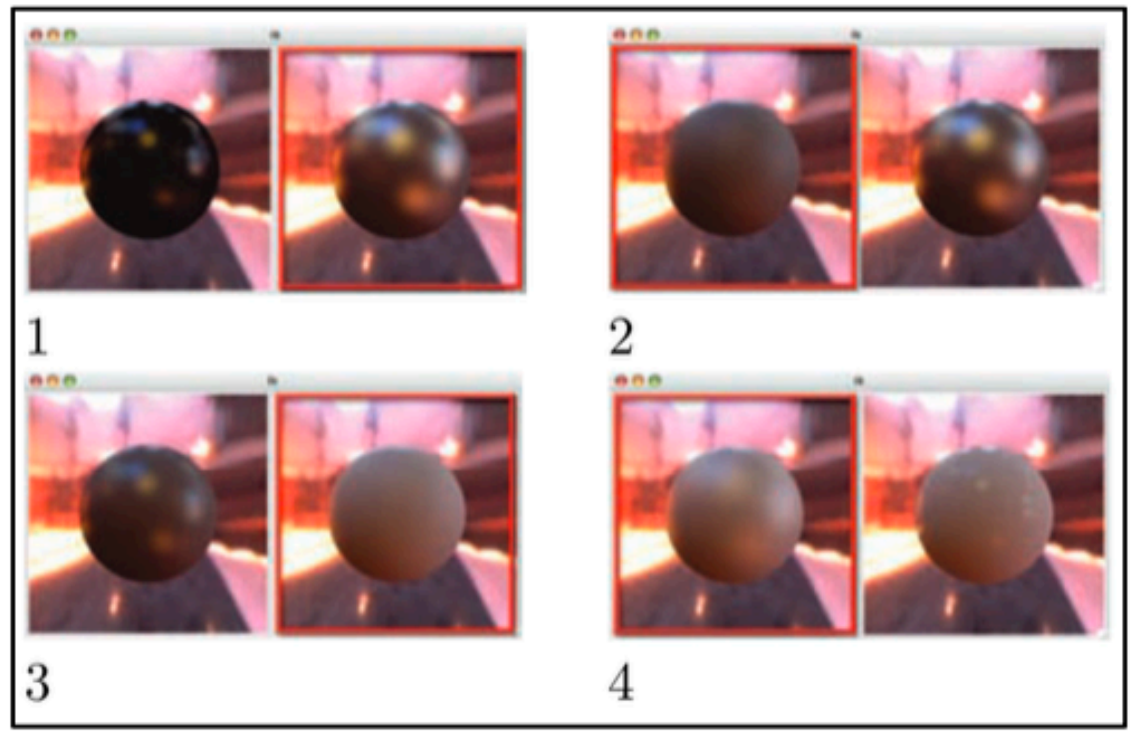
[Marks et al. 1997]



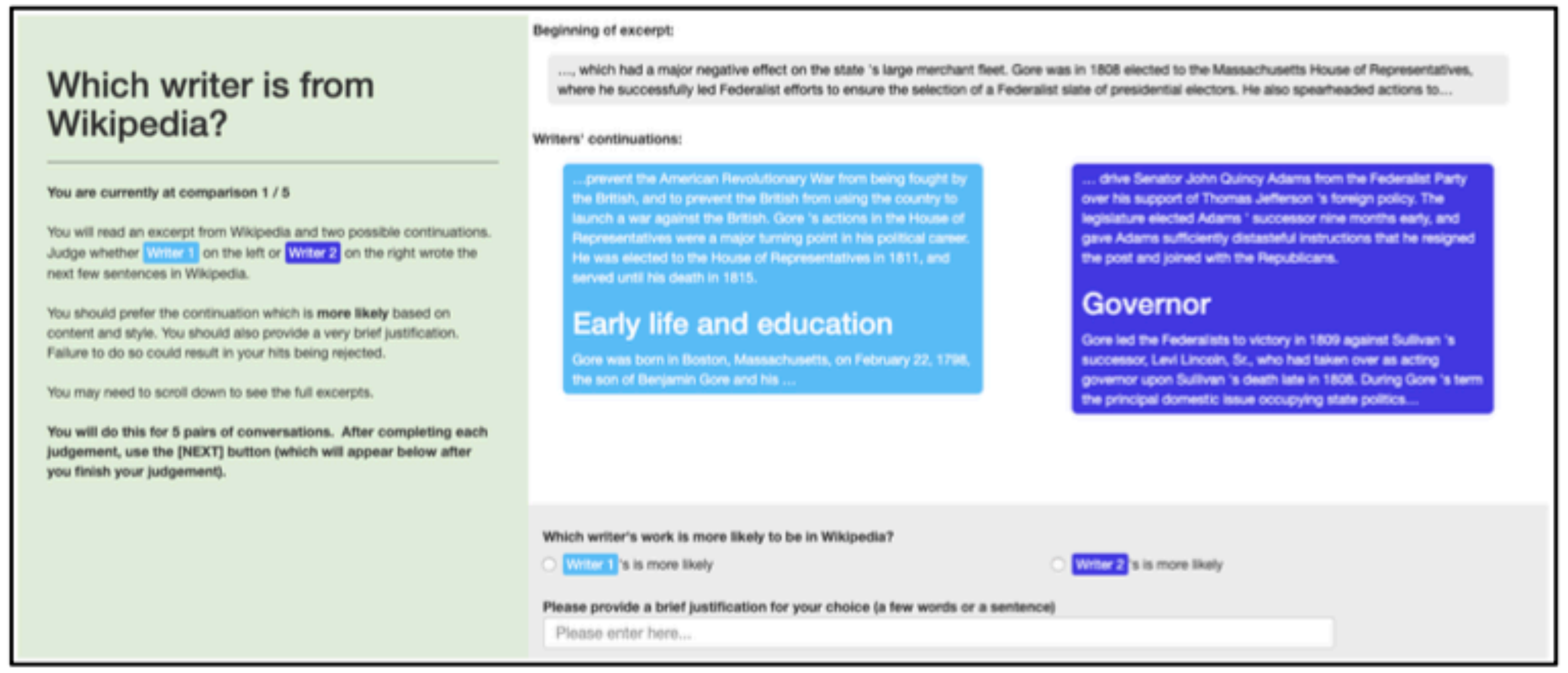
[Wilber et al. 2014]

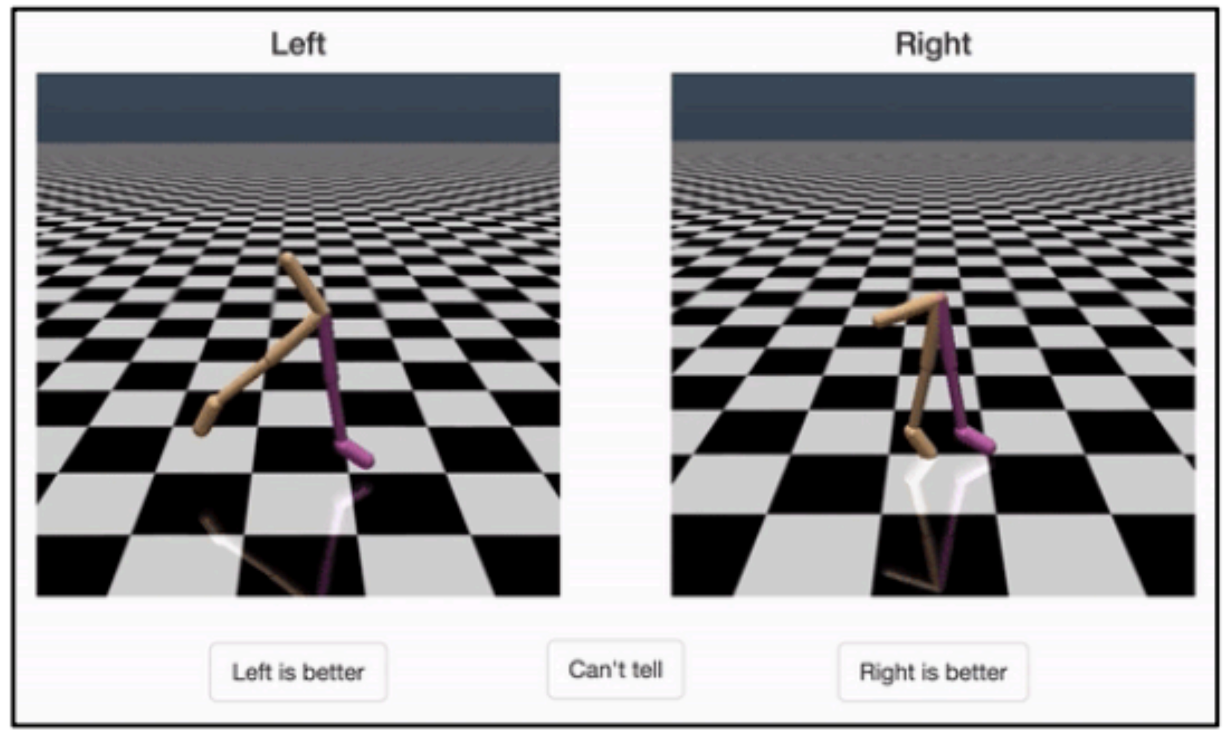


[Brochu et al. 2007]

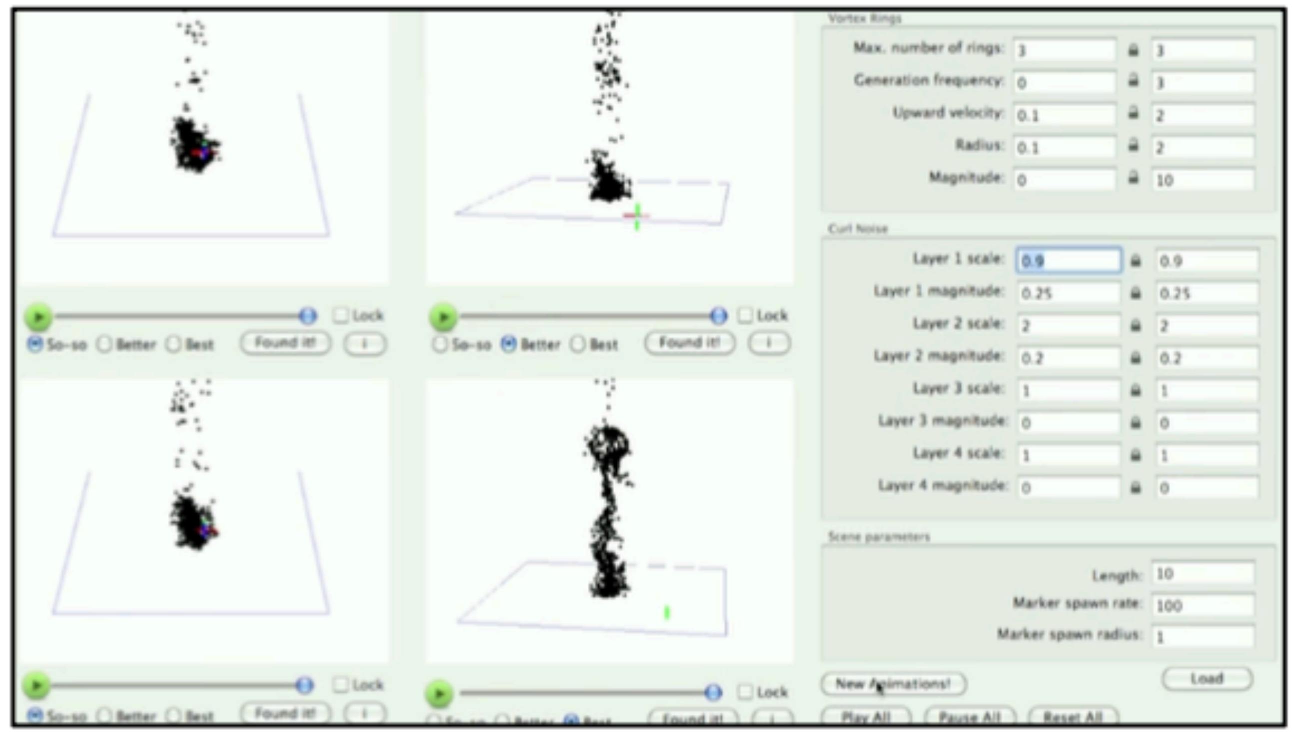


[Welleck et al. 2020]



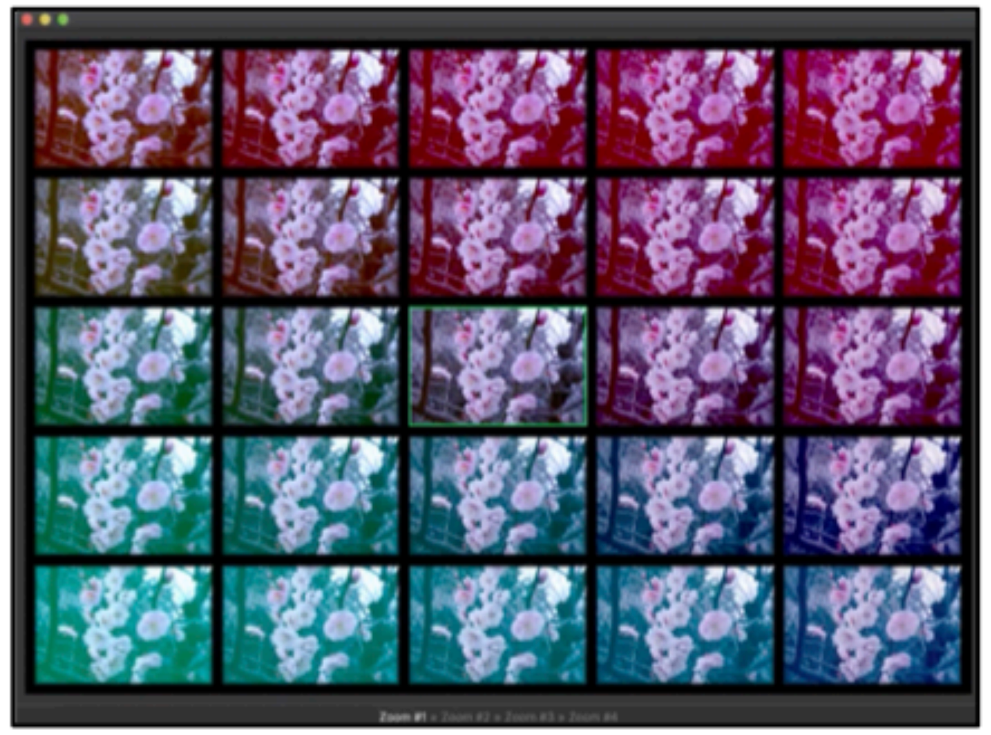


[Christiano et al. 2017]



[Brochu et al. 2007]

[Koyama et al. 2014]



[Koyama et al. 2020]

Opinion Measurement UIs for HITL Systems

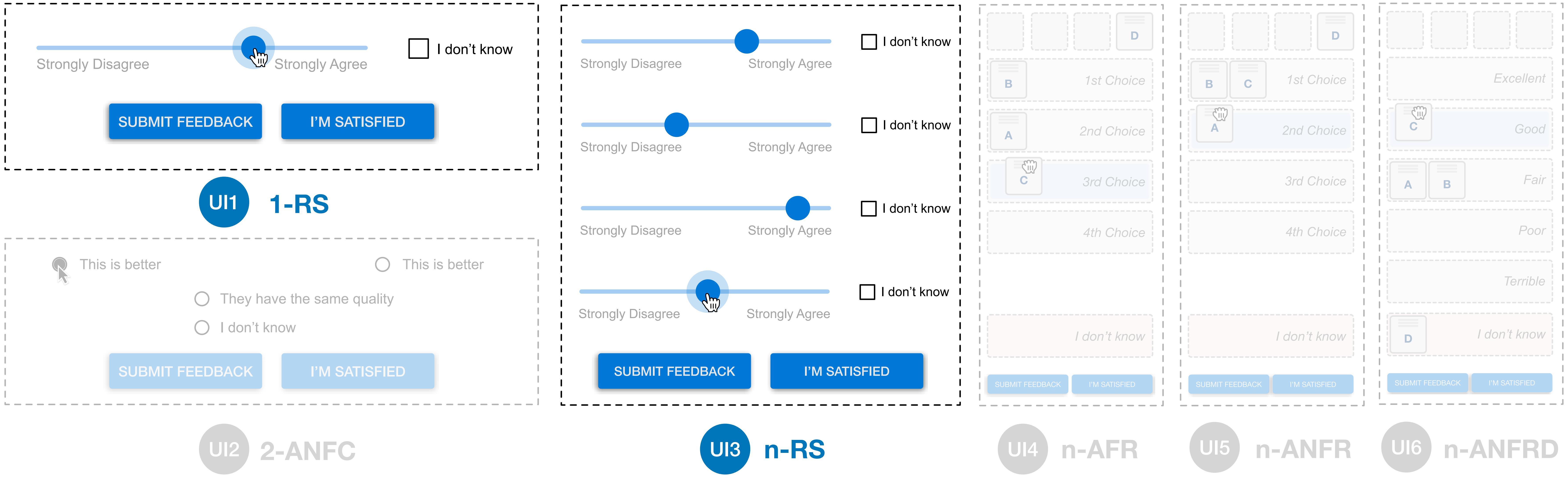
1-RS, n-RS: single (or n) rating scale(s), bipolar

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

	Alternative Options (n = 1, 2, 3, …)		
Feedback Type	**Pointwise** (n=1)	**Pairwise** (n=2)	**Listwise** (n>2)
Utility	1-RS (***UI1***)	2-RS	n-RS (***UI3***)
Preferential *(strict)*		2-AFC	n-AFR (***UI4***)
Preferential *(weak)*		2-ANFC (***UI2***)	n-ANFR (***UI5***)
Hybrid *(strict)*			n-AFRD
Hybrid *(weak)*			n-ANFRD (***UI6***)



Opinion Measurement UIs for HITL Systems

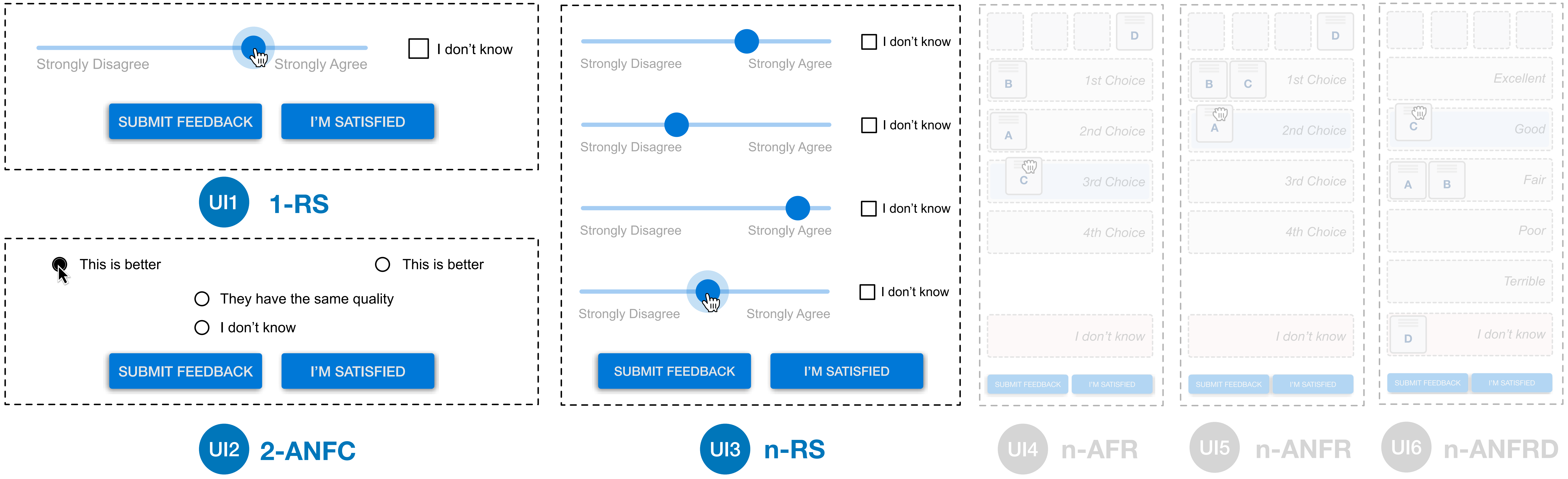
1-RS, n-RS: single (or n) rating scale(s), bipolar

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

	Alternative Options (n = 1, 2, 3, ...)		
Feedback Type	**Pointwise** (n=1)	**Pairwise** (n=2)	**Listwise** (n>2)
Utility	1-RS (***UI1***)	2-RS	n-RS (***UI3***)
Preferential *(strict)*		2-AFC	n-AFR (***UI4***)
Preferential *(weak)*		2-ANFC (***UI2***)	n-ANFR (***UI5***)
Hybrid *(strict)*			n-AFRD
Hybrid *(weak)*			n-ANFRD (***UI6***)



Opinion Measurement UIs for HITL Systems

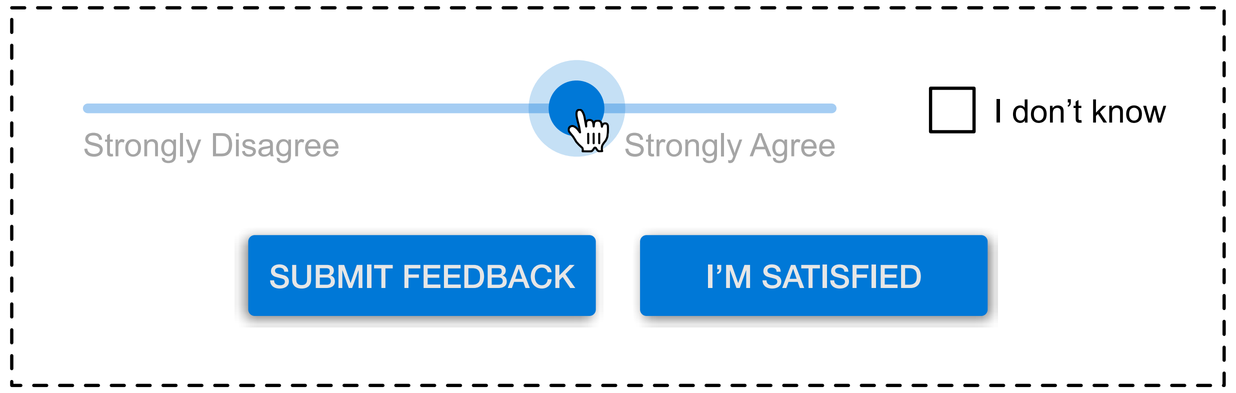
1-RS, n-RS: single (or n) rating scale(s), bipolar

2-ANFC: two alternative non-forced choice

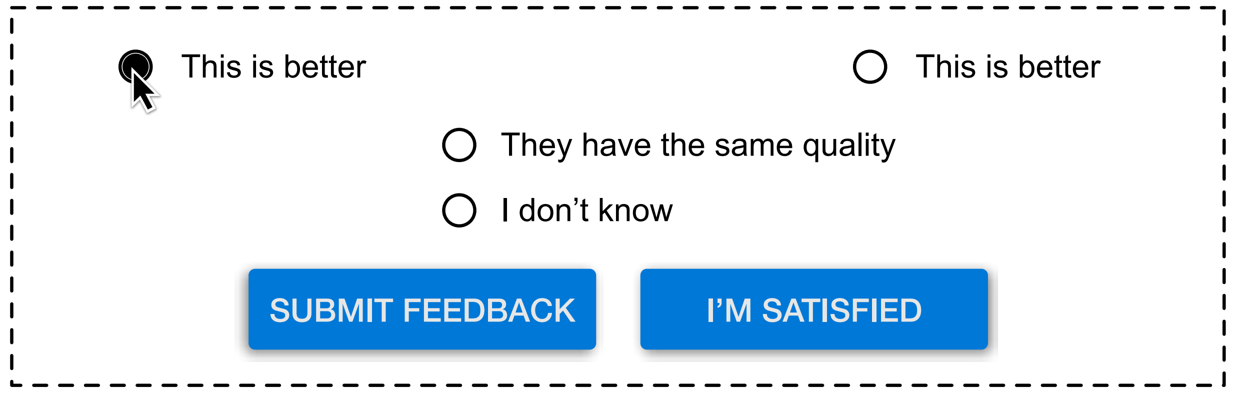
n-AFR, n-ANFR: n-alternative (non-)forced ranking

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

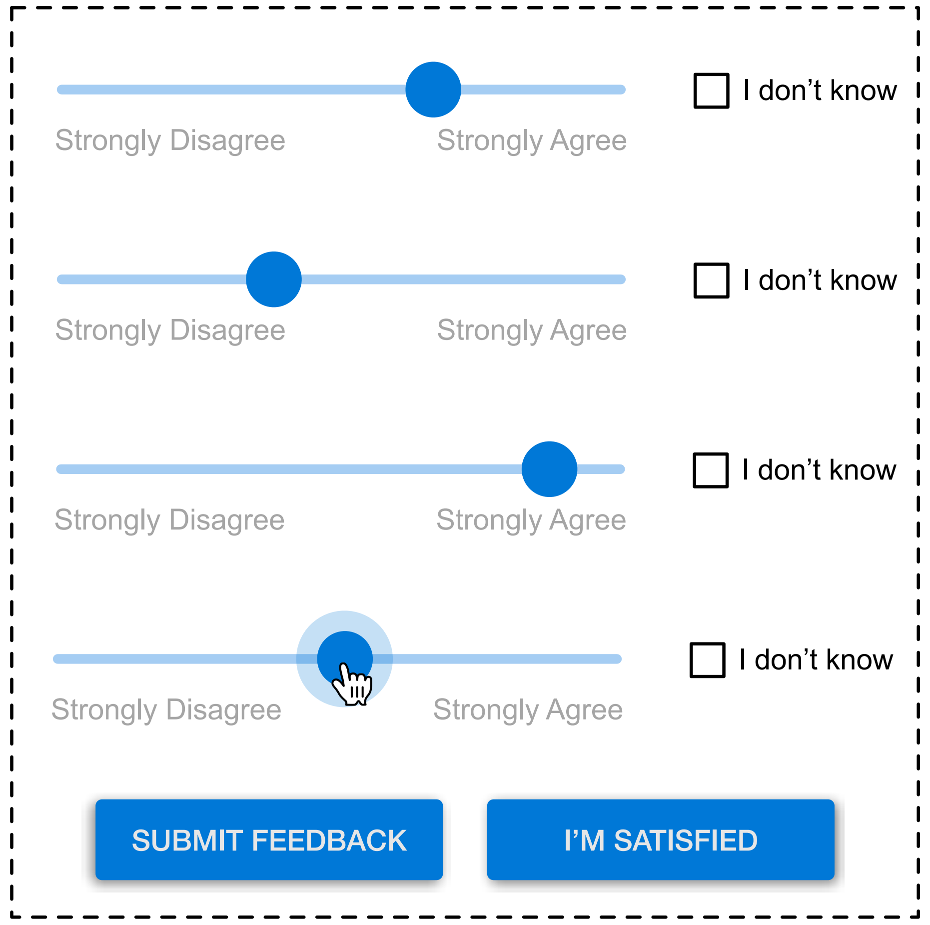
	Alternative Options (n = 1, 2, 3, …)		
Feedback Type	**Pointwise** (n=1)	**Pairwise** (n=2)	**Listwise** (n>2)
Utility	1-RS (***UI1***)	2-RS	n-RS (***UI3***)
Preferential *(strict)*		2-AFC	n-AFR (***UI4***)
Preferential *(weak)*		2-ANFC (***UI2***)	n-ANFR (***UI5***)
Hybrid *(strict)*			n-AFRD
Hybrid *(weak)*			n-ANFRD (***UI6***)



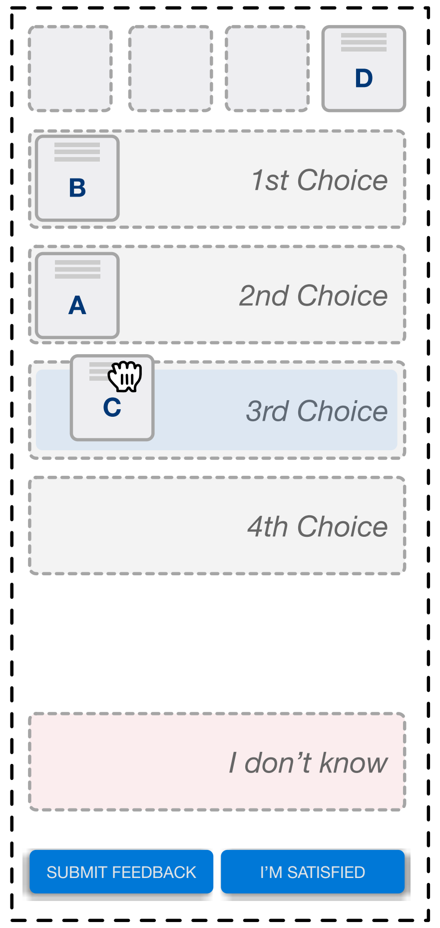
UI1 1-RS



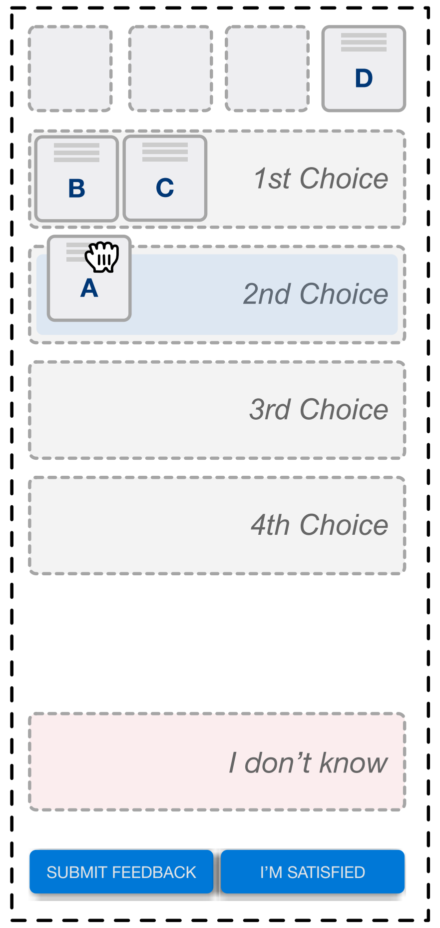
UI2 2-ANFC



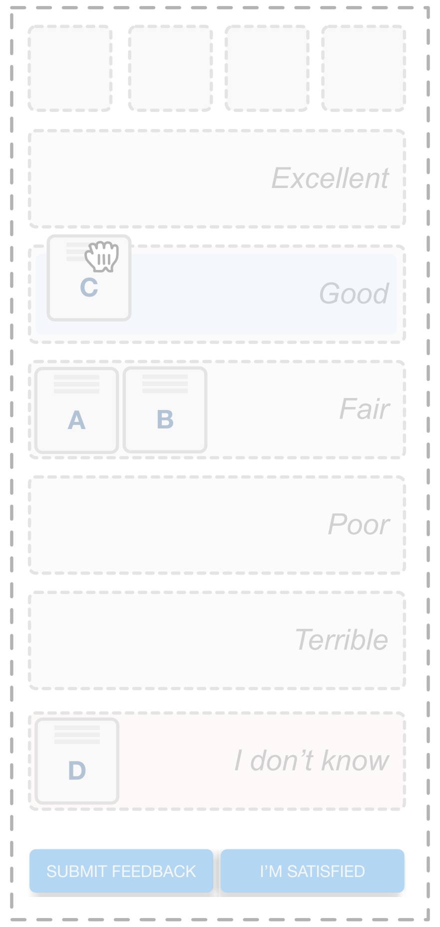
UI3 n-RS



UI4 n-AFR



UI5 n-ANFR



UI6 n-ANFRD

Opinion Measurement UIs for HITL Systems

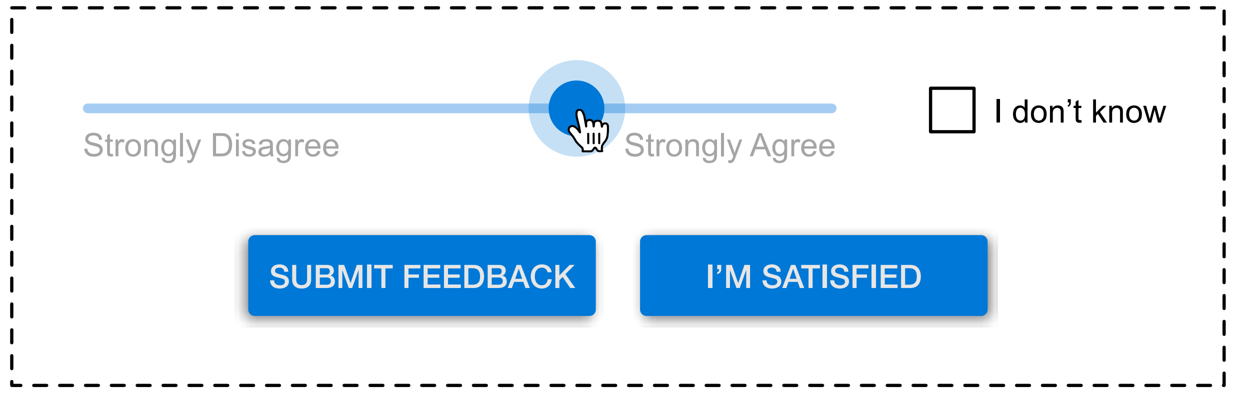
1-RS, n-RS: single (or n) rating scale(s), bipolar

2-ANFC: two alternative non-forced choice

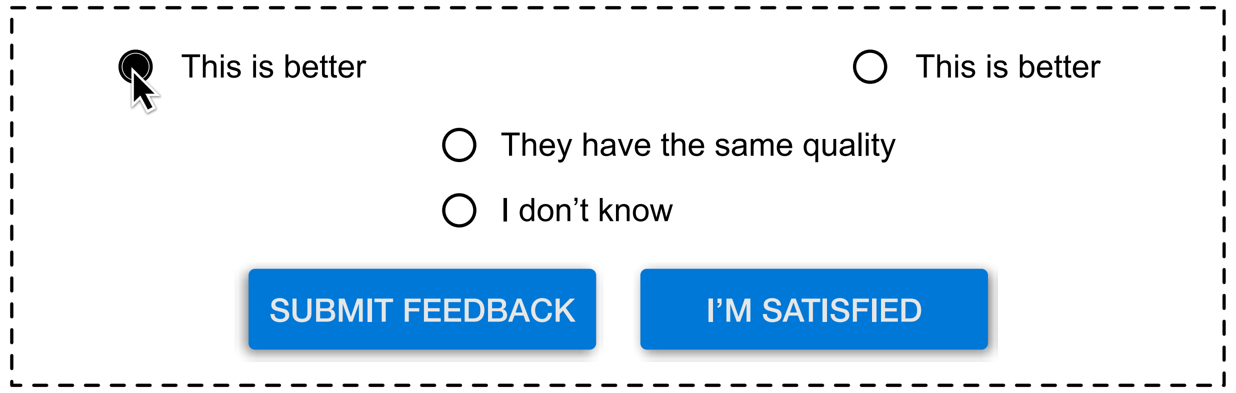
n-AFR, n-ANFR: n-alternative (non-)forced ranking

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

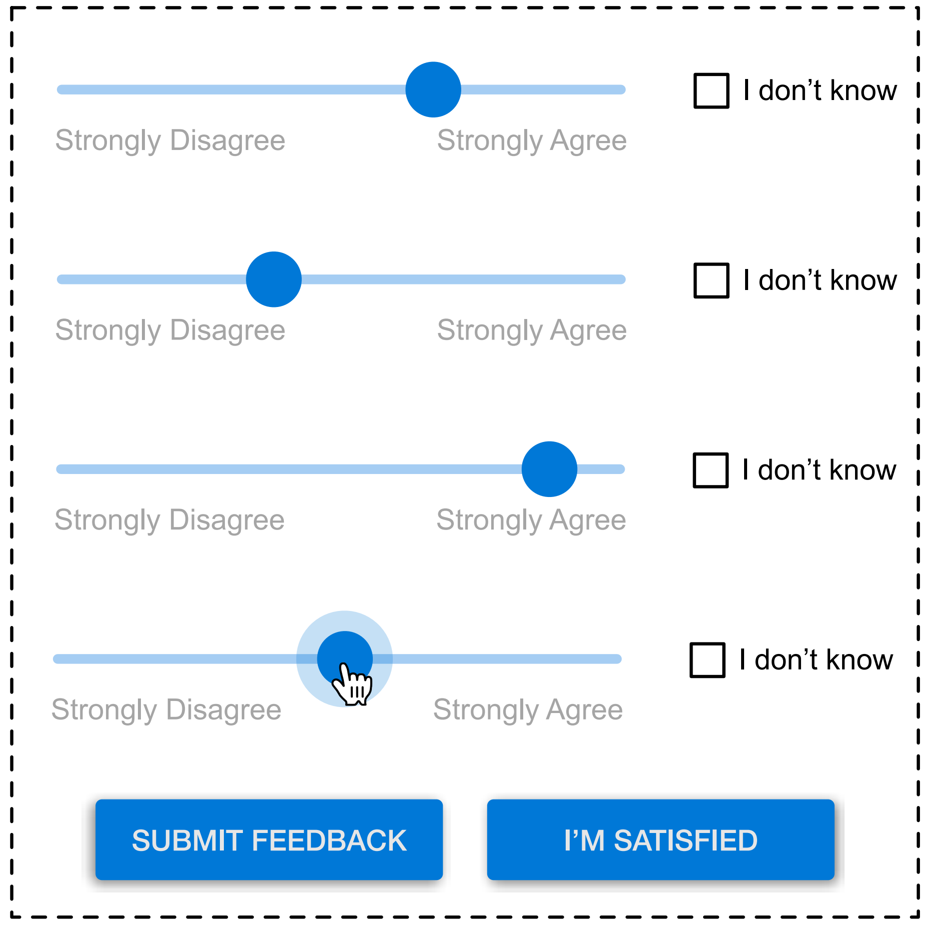
	Alternative Options (n = 1, 2, 3, ...)		
Feedback Type	**Pointwise** (n=1)	**Pairwise** (n=2)	**Listwise** (n>2)
Utility	1-RS (***UI1***)	2-RS	n-RS (***UI3***)
Preferential *(strict)*		2-AFC	n-AFR (***UI4***)
Preferential *(weak)*		2-ANFC (***UI2***)	n-ANFR (***UI5***)
Hybrid *(strict)*			n-AFRD
Hybrid *(weak)*			n-ANFRD (***UI6***)



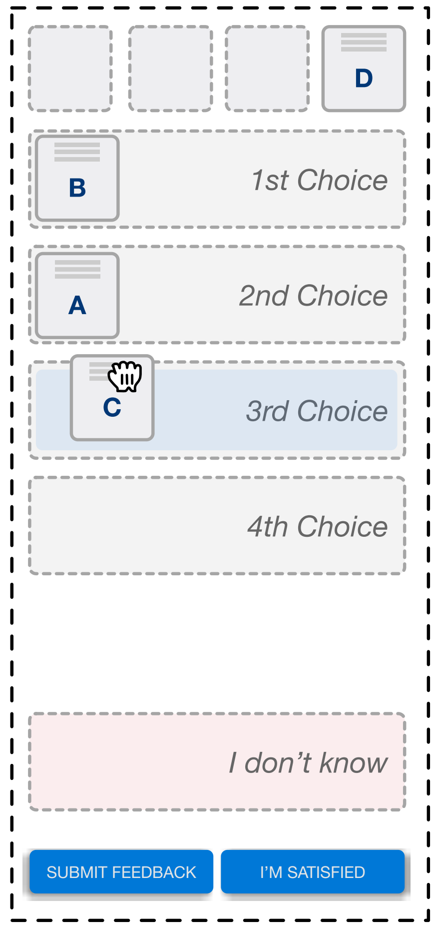
UI1 **1-RS**



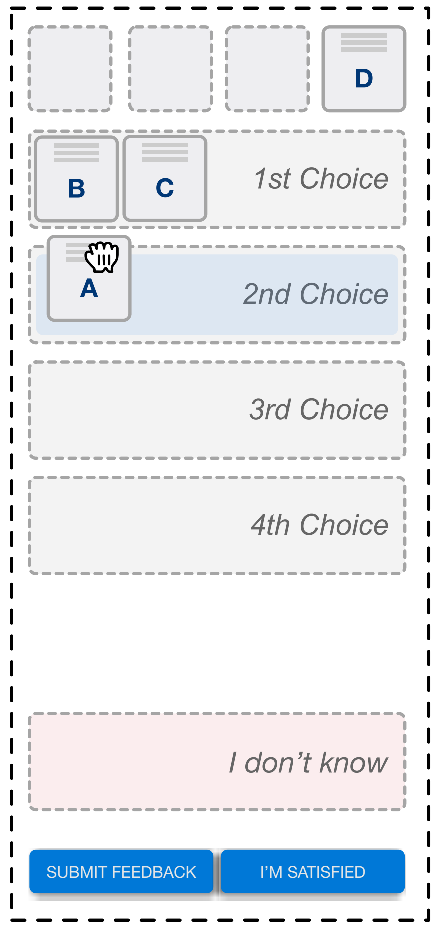
UI2 **2-ANFC**



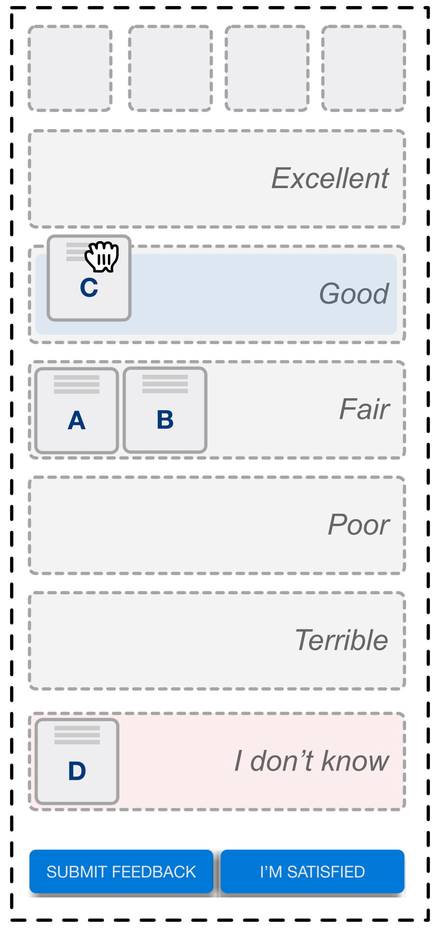
UI3 **n-RS**



UI4 **n-AFR**



UI5 **n-ANFR**

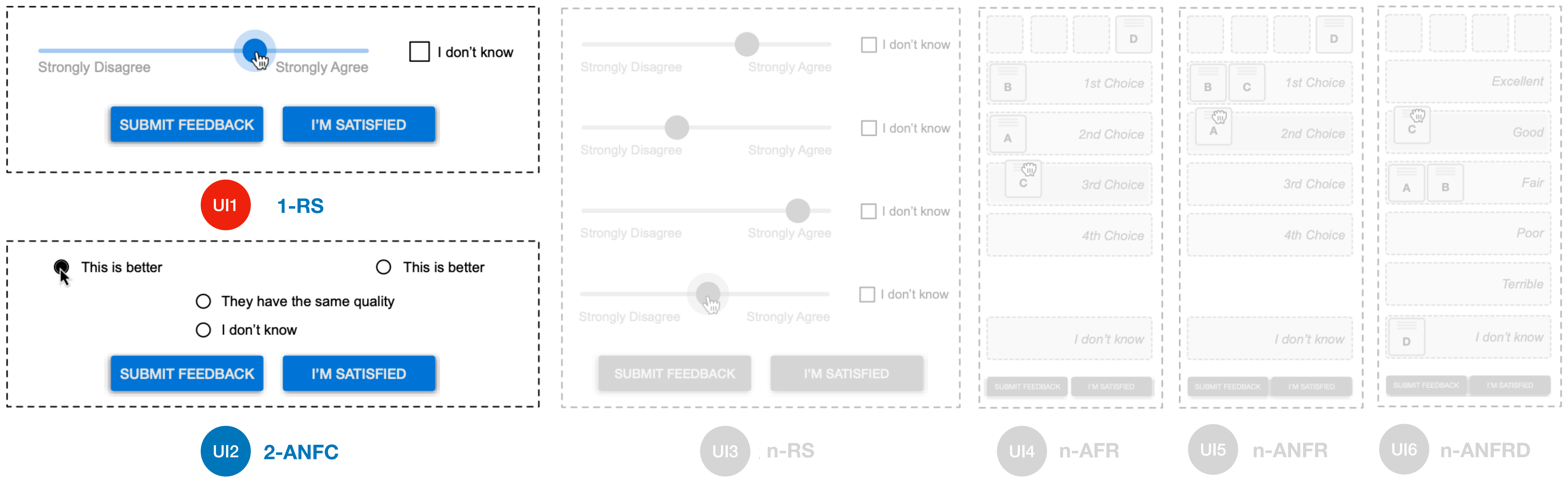


UI6 **n-ANFRD**

Hypotheses for UI Ranking

H1 Baseline

- UI1 < UI2



	Alternative Options (n = 1, 2, 3, …)		
Feedback Type	**Pointwise** (n=1)	**Pairwise** (n=2)	**Listwise** (n>2)
Utility	1-RS (***UI1***)	2-RS	n-RS (***UI3***)
Preferential *(strict)*		2-AFC	n-AFR (***UI4***)
Preferential *(weak)*		2-ANFC (***UI2***)	n-ANFR (***UI5***)
Hybrid *(strict)*			n-AFRD
Hybrid *(weak)*			n-ANFRD (***UI6***)

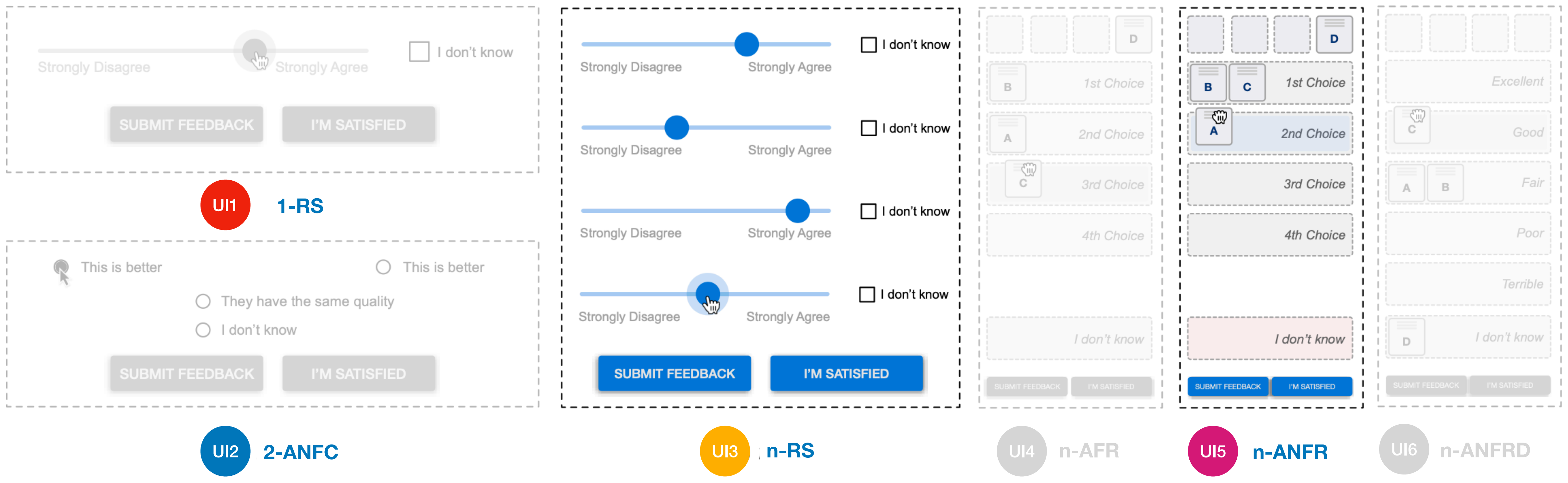
Hypotheses for UI Ranking

H1 Baseline

- UI1 < UI2

H2 Impact of Listwise UI

- UI1 < UI3
- UI2 < UI5



	Alternative Options (n = 1, 2, 3, …)		
Feedback Type	**Pointwise** (n=1)	**Pairwise** (n=2)	**Listwise** (n>2)
Utility	1-RS (***UI1***)	2-RS	n-RS (***UI3***)
Preferential *(strict)*		2-AFC	n-AFR (***UI4***)
Preferential *(weak)*		2-ANFC (***UI2***)	n-ANFR (***UI5***)
Hybrid *(strict)*			n-AFRD
Hybrid *(weak)*			n-ANFRD (***UI6***)

Hypotheses for UI Ranking

H1 Baseline

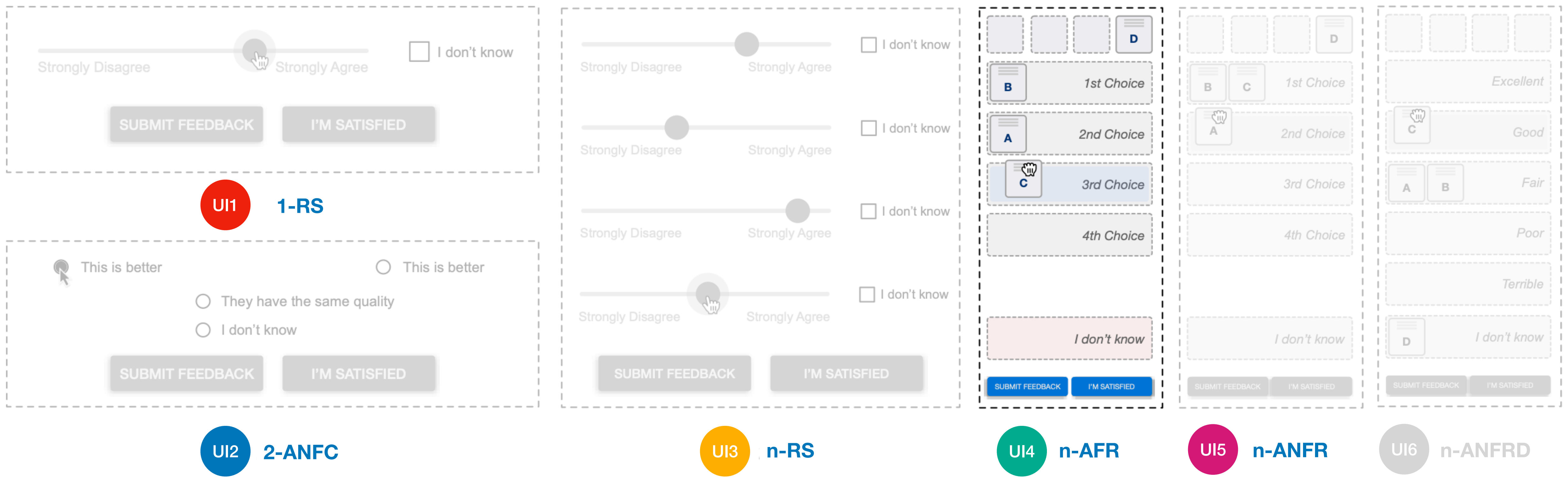
- UI1 < UI2

H2 Impact of Listwise UI

- UI1 < UI3
- UI2 < UI5

H3 Impact of Listwise Design Variation

- UI3 < UI4 < UI5



	Alternative Options (n = 1, 2, 3, …)		
Feedback Type	**Pointwise** (n=1)	**Pairwise** (n=2)	**Listwise** (n>2)
Utility	1-RS (***UI1***)	2-RS	n-RS (***UI3***)
Preferential *(strict)*		2-AFC	n-AFR (***UI4***)
Preferential *(weak)*		2-ANFC (***UI2***)	n-ANFR (***UI5***)
Hybrid *(strict)*			n-AFRD
Hybrid *(weak)*			n-ANFRD (***UI6***)

Hypotheses for UI Ranking

H1 Baseline

- UI1 < UI2

H2 Impact of Listwise UI

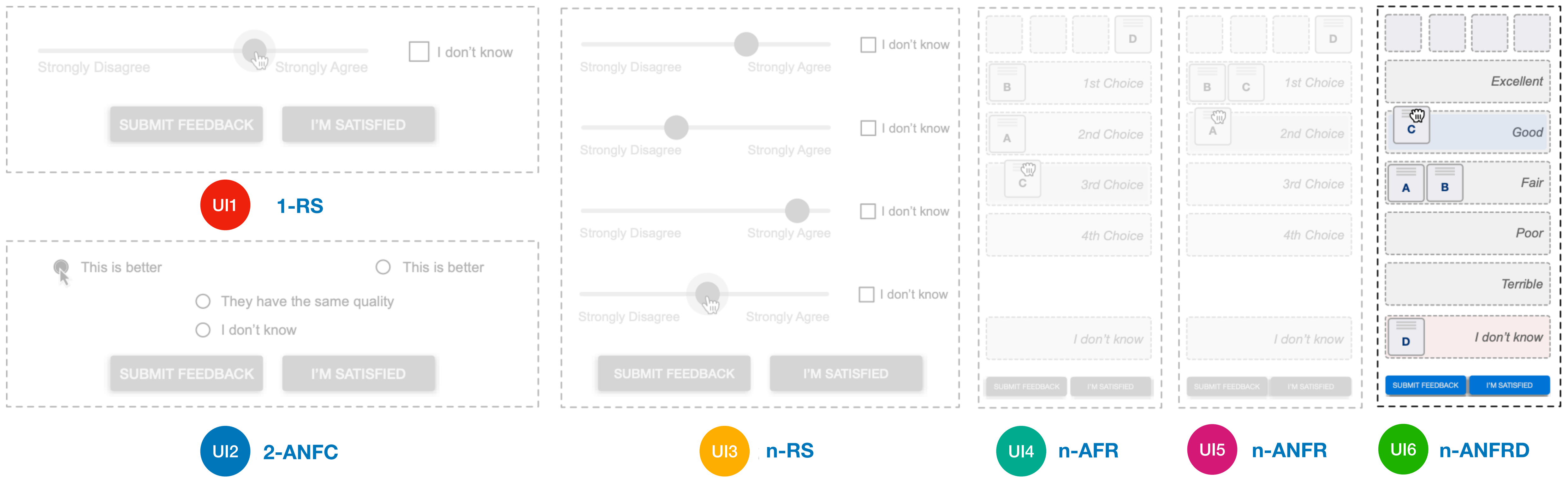
- UI1 < UI3
- UI2 < UI5

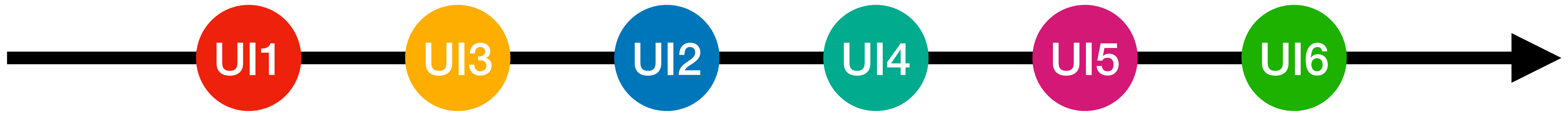
H3 Impact of Listwise Design Variation

- UI3 < UI4 < UI5

H4 Impact of Hybrid UI

- UI3 < UI2 < UI4 < UI5 < UI6





	Alternative Options (n = 1, 2, 3, …)		
Feedback Type	**Pointwise** (n=1)	**Pairwise** (n=2)	**Listwise** (n>2)
Utility	1-RS (***UI1***)	2-RS	n-RS (***UI3***)
Preferential *(strict)*		2-AFC	n-AFR (***UI4***)
Preferential *(weak)*		2-ANFC (***UI2***)	n-ANFR (***UI5***)
Hybrid *(strict)*			n-AFRD
Hybrid *(weak)*			n-ANFRD (***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:

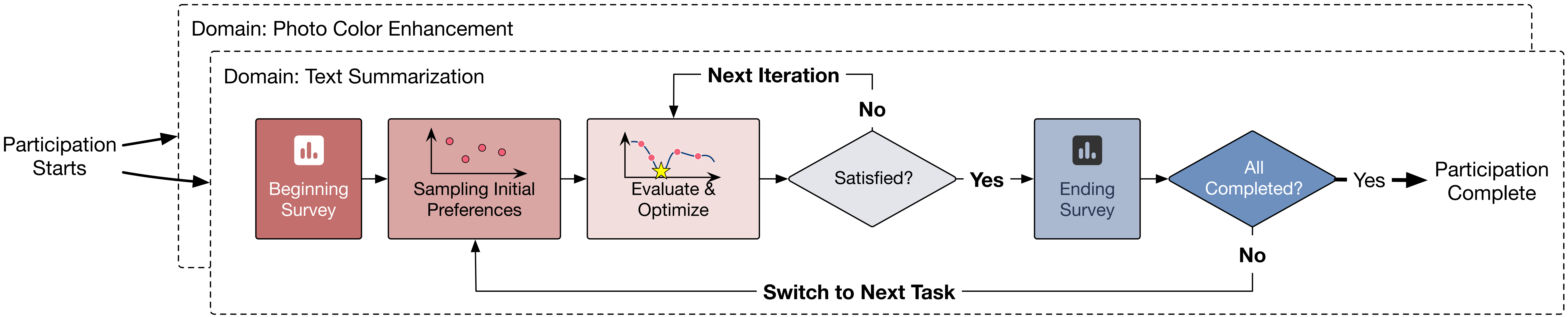
Fill beginning survey, providing feedback to the AI results; inspect improved results; loop until satisfaction, and fill ending survey

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)

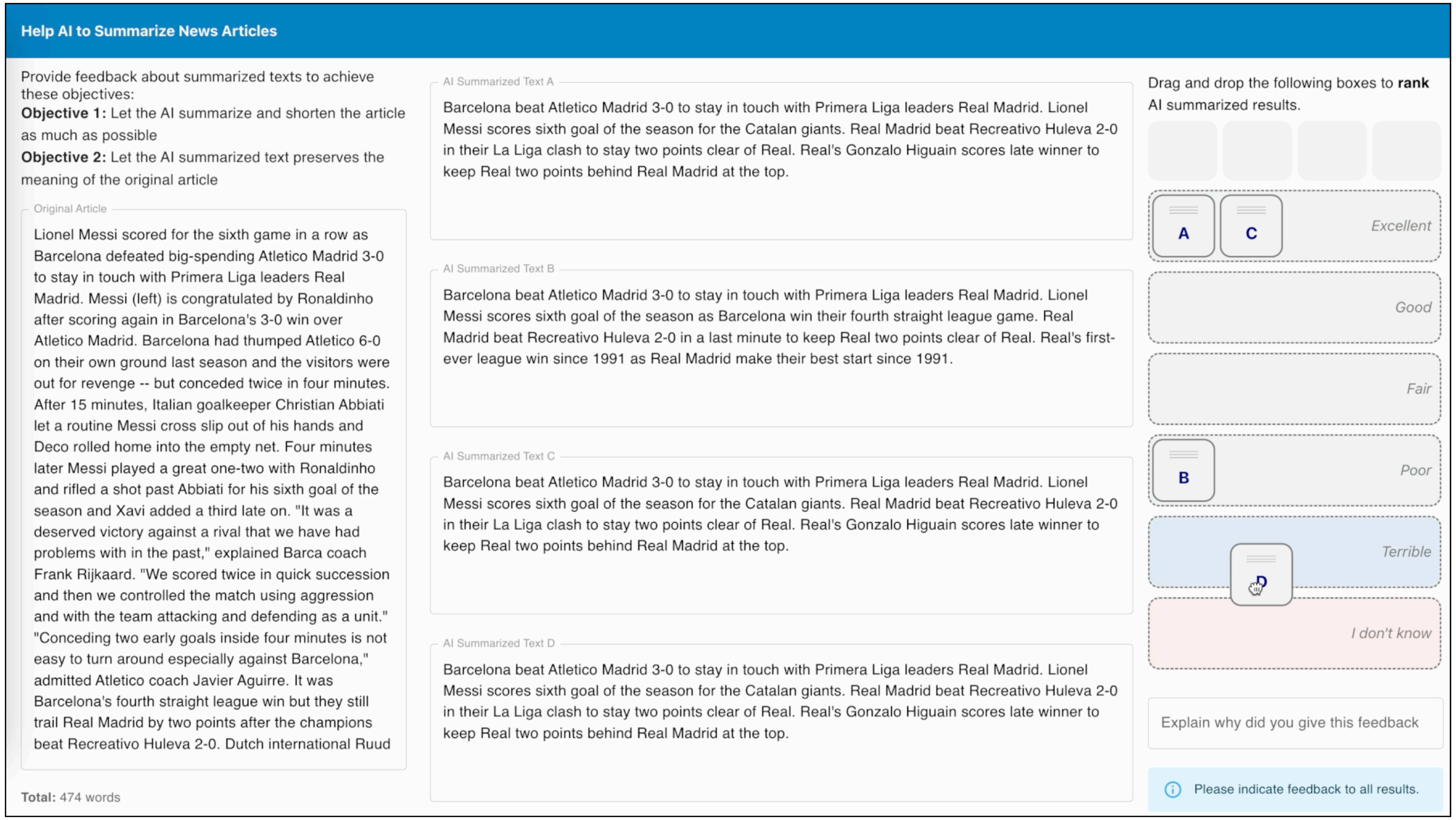


Apparatus: Text Summarization

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

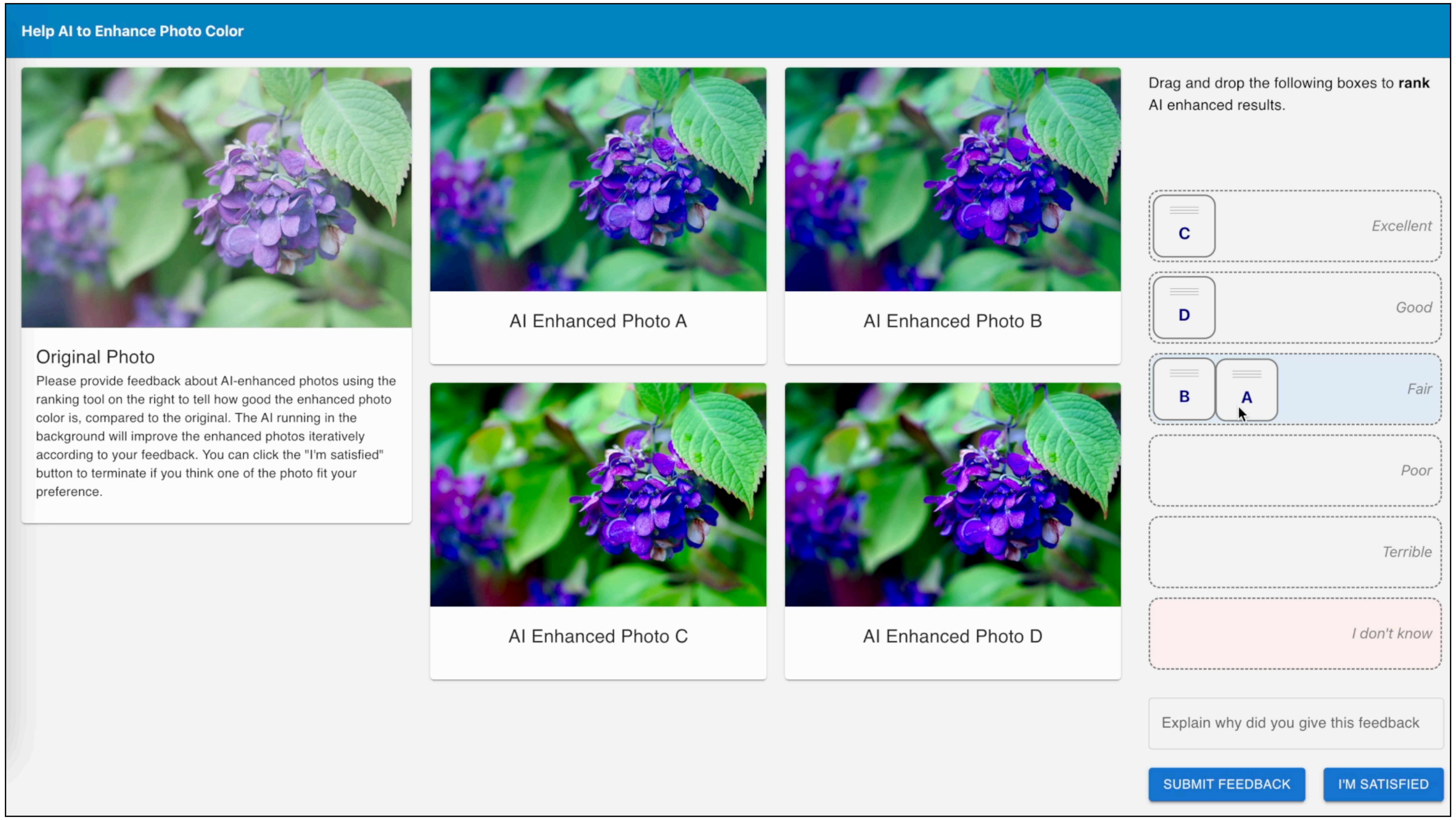


Apparatus: Photo Color Enhancement

Photo color enhancer [Koyama et al. 2016, 2017, 2020]

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

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.

Example outcomes

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 $f([\mathrm{x}, \mathrm{x}']) = g(\mathrm{x}') - g(\mathrm{x})$ that defines a **preference function**

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

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

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

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

$$[\mathbf{x}_{j+1}, \mathbf{x}_{j+1}'] = [\operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \int_{\mathcal{X}} \pi_{\tilde{f}}([\mathbf{x}, \mathbf{x}']; D_j) \mathrm{d}\mathbf{x}', \operatorname*{argmax}_{\mathbf{x}_*' \in \mathcal{X}} \mathbf{V}[\sigma(f_*) | [\mathbf{x}_*, \mathbf{x}_*'], \mathcal{D}_j, \mathbf{x}_* = \mathbf{x}_{j+1}]]$$

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

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

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:

$$V(\mathbf{x}, \mathbf{x}') = \mathbf{E}_j \left[\underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \, \mathbf{E}_{j+1}[f(\mathbf{x})] - \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \, \mathbf{E}_j[f(\mathbf{x})] \right]$$

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:

$$V(\mathbf{x}, \mathbf{x}') = \mathbf{E}_j \left[\operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \mathbf{E}_{j+1}[f(\mathbf{x})] - \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \mathbf{E}_j[f(\mathbf{x})] \right]$$

Maximizing this AF is equivalent to

$$\operatorname*{argmax}_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \mathrm{EUBO}(\mathbf{x}, \mathbf{x}') = \mathbf{E}_j[\max\{f(\mathbf{x}), f(\mathbf{x}')\}] \subseteq \operatorname*{argmax}_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} V(\mathbf{x}, \mathbf{x}')$$

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

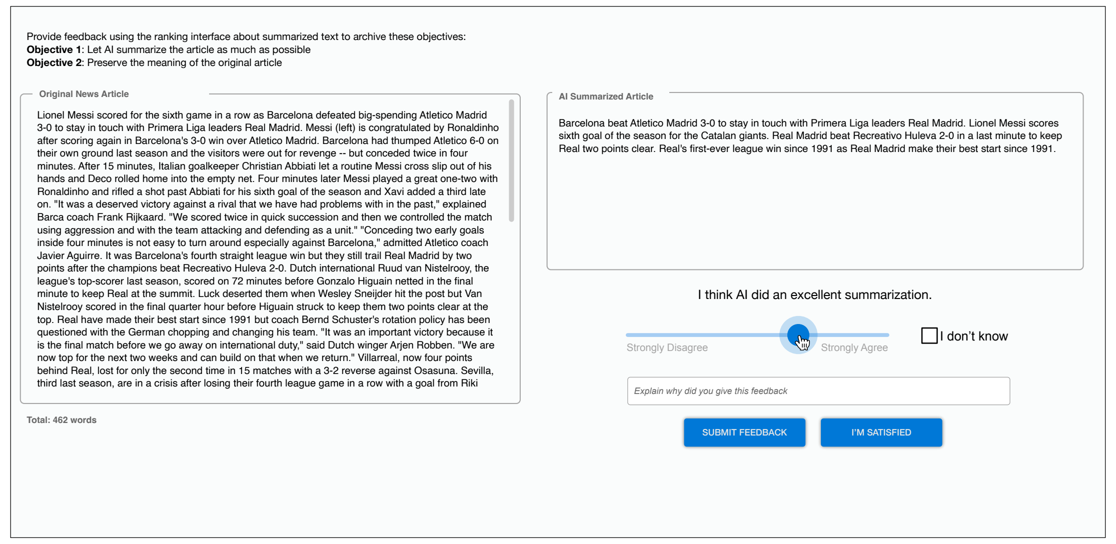
- Objective outcome quality measures
  - BLEU, ROUGE; MSE chan diff in HSV and YUV space
- Optimizer measures
  - Posterior mean of the estimated ranking utility (*Latent Utility*)

User performance

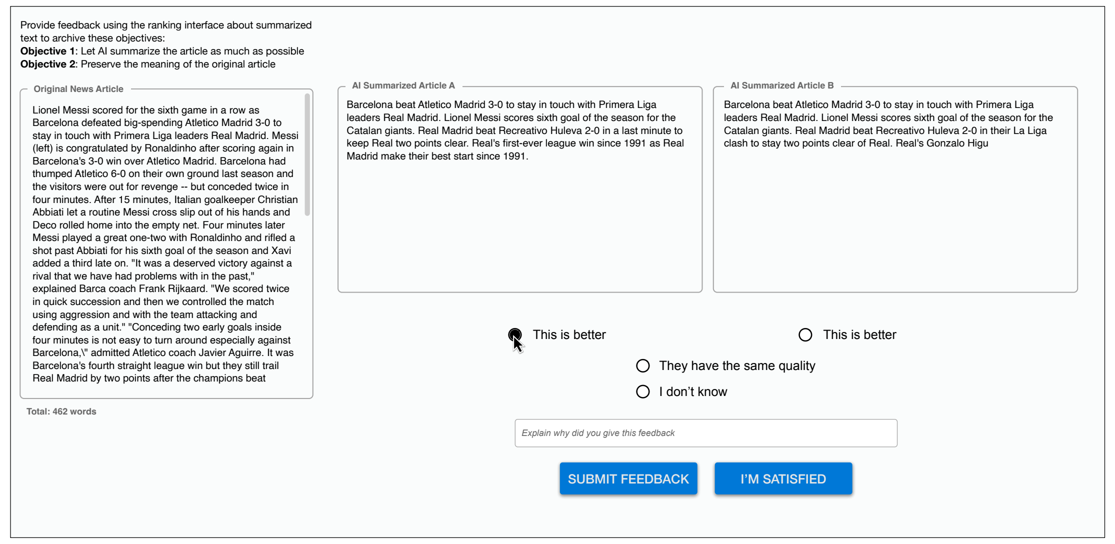
- Interaction behavior measures
  - decision time, iterations, incomplete/indifference preference, ranking interactions
- User’s actual input rating/ranking utility for the machine outcomes (*Direct Utility*)
- subjective satisfaction

Model formula: `perf ~ UI * progress + (1|participant) + (1|task)`

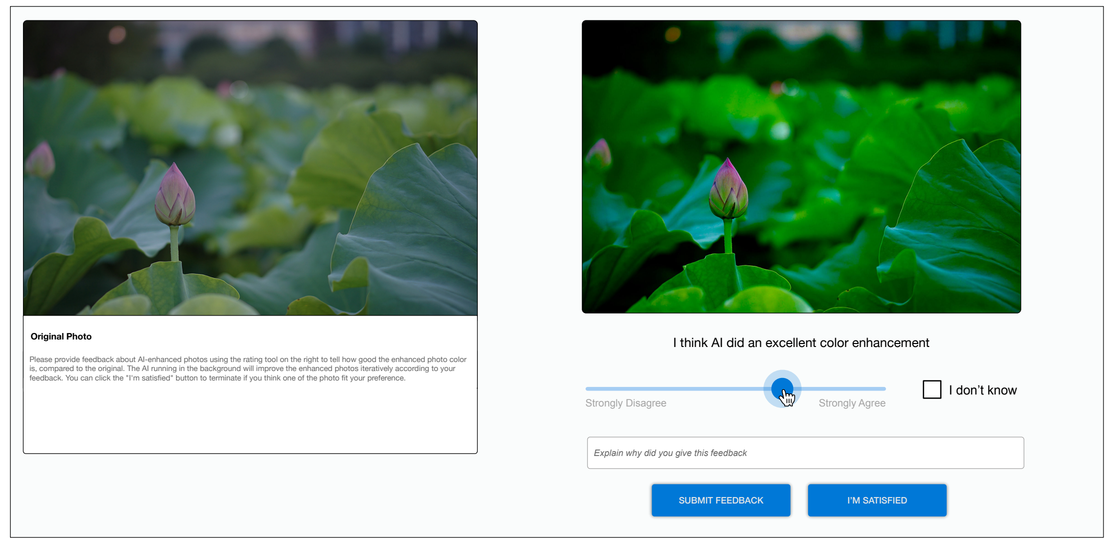
Apparatus: User Interfaces



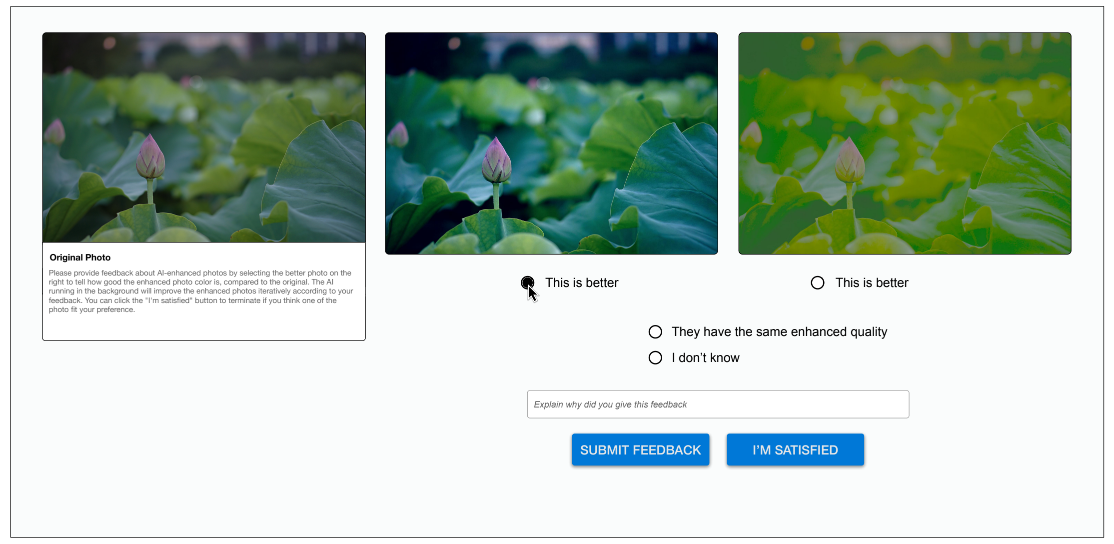
UI1) 1-RS



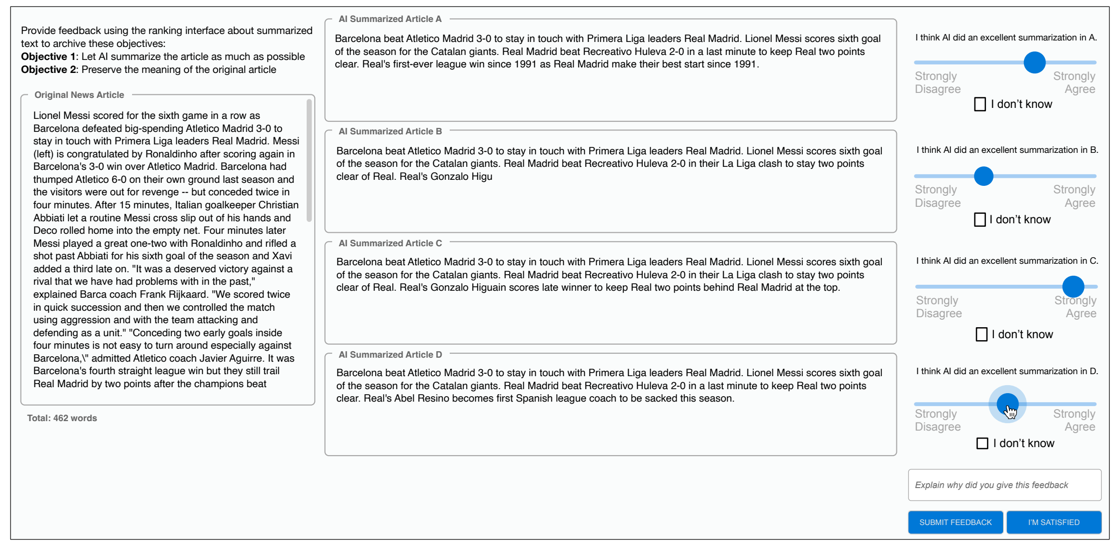
UI2) 2-ANFC



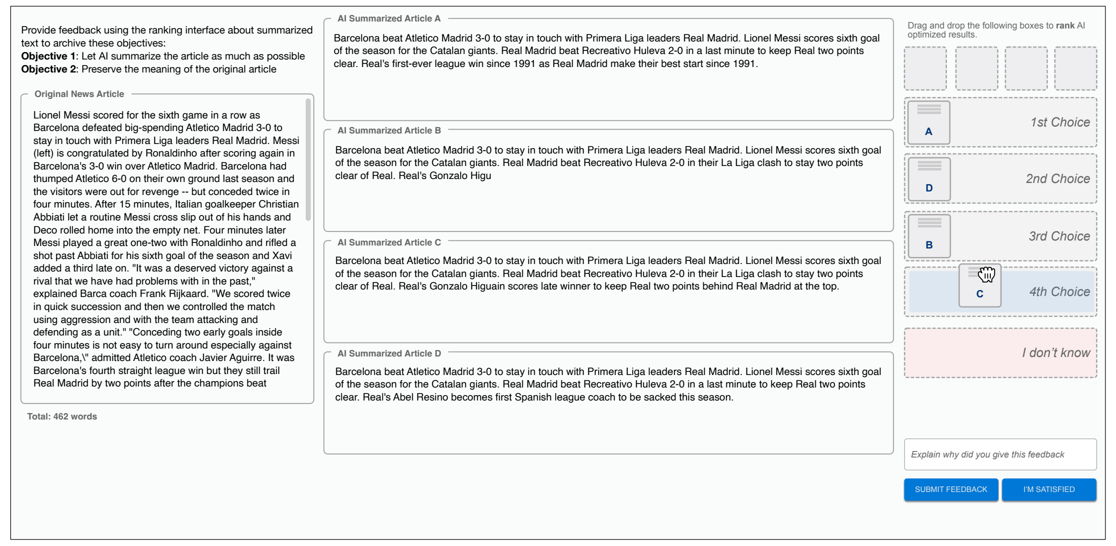
UI1) 1-RS



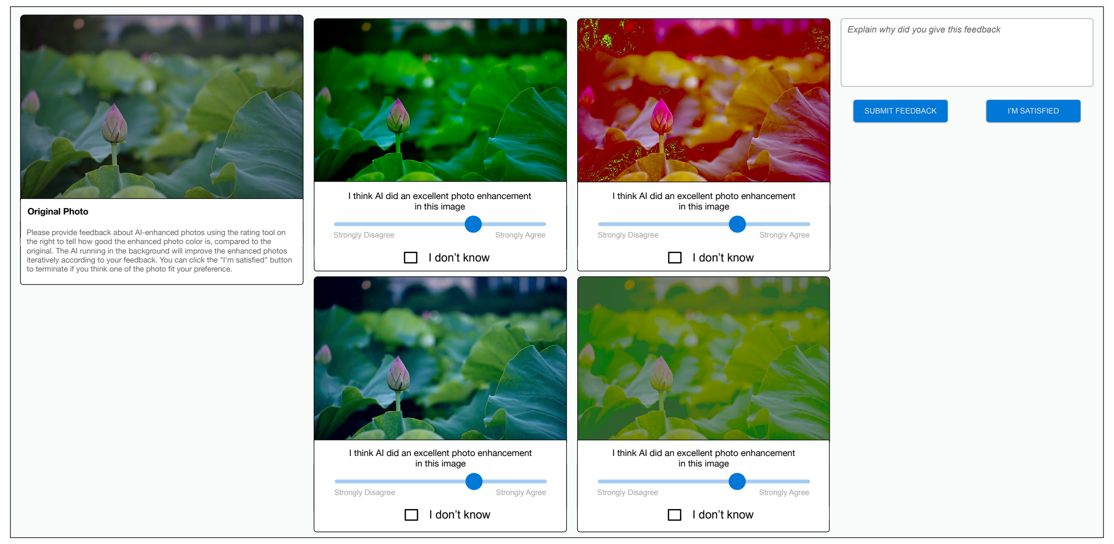
UI2) 2-ANFC



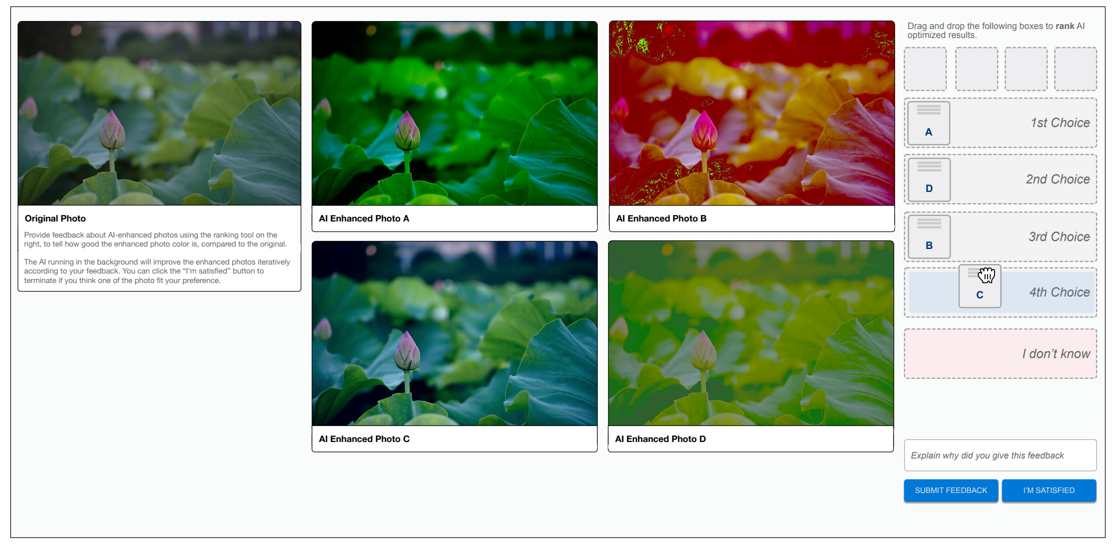
UI3) *n*-RS



UI4) *n*-AFR



UI3) *n*-RS

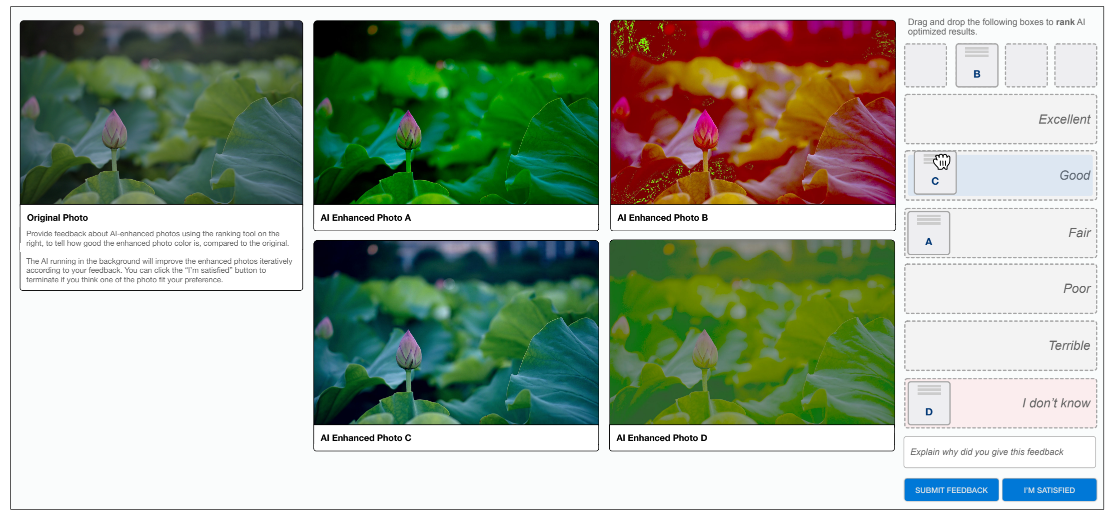


UI4) *n*-AFR

UI5) *n*-ANFR

UI6) *n*-ANFRD

UI5) *n*-ANFR



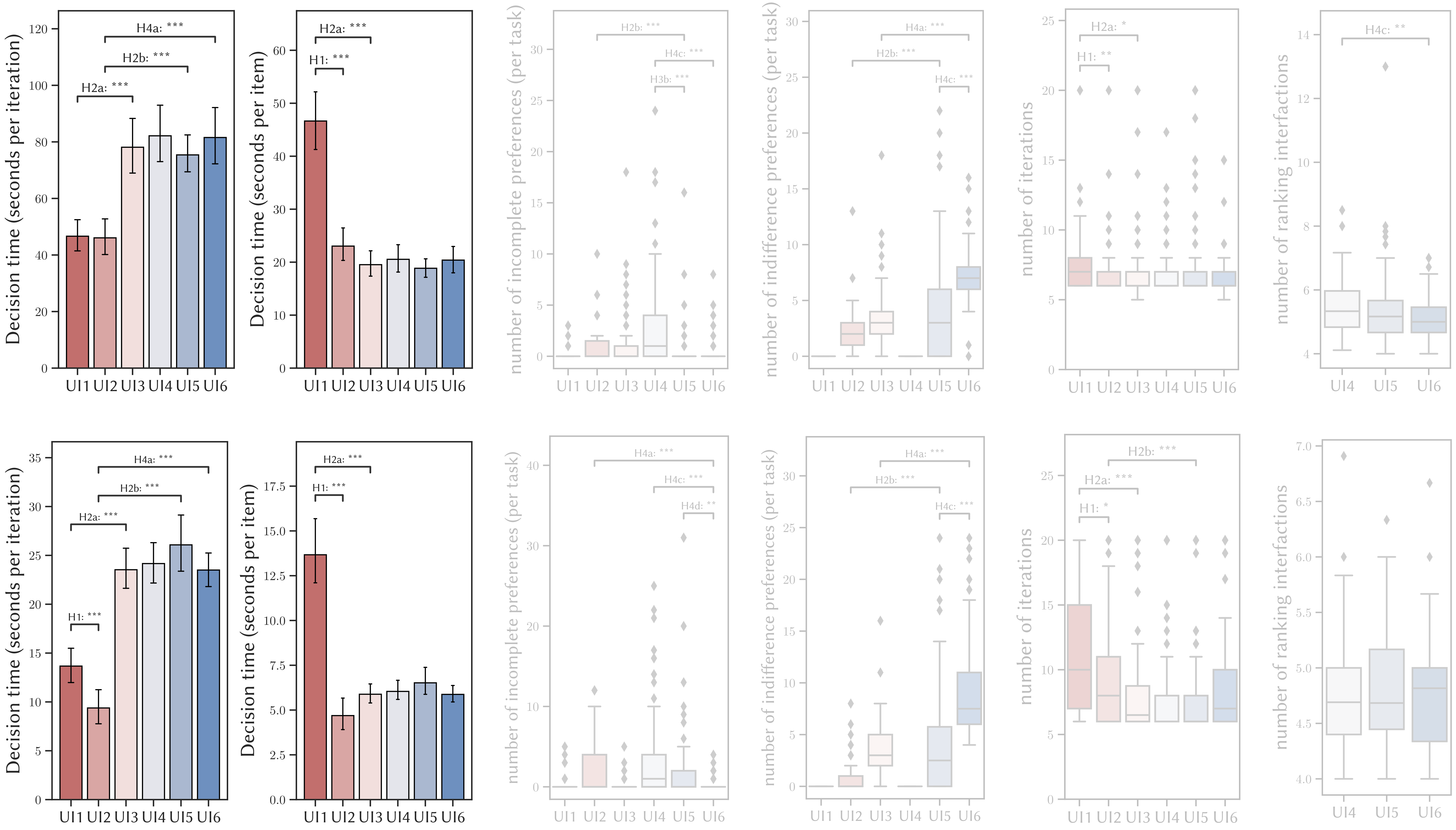
UI6) *n*-ANFRD

User Performance: Interaction Behavior

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

Text Summarization Task

Photo Color Enhancement Task



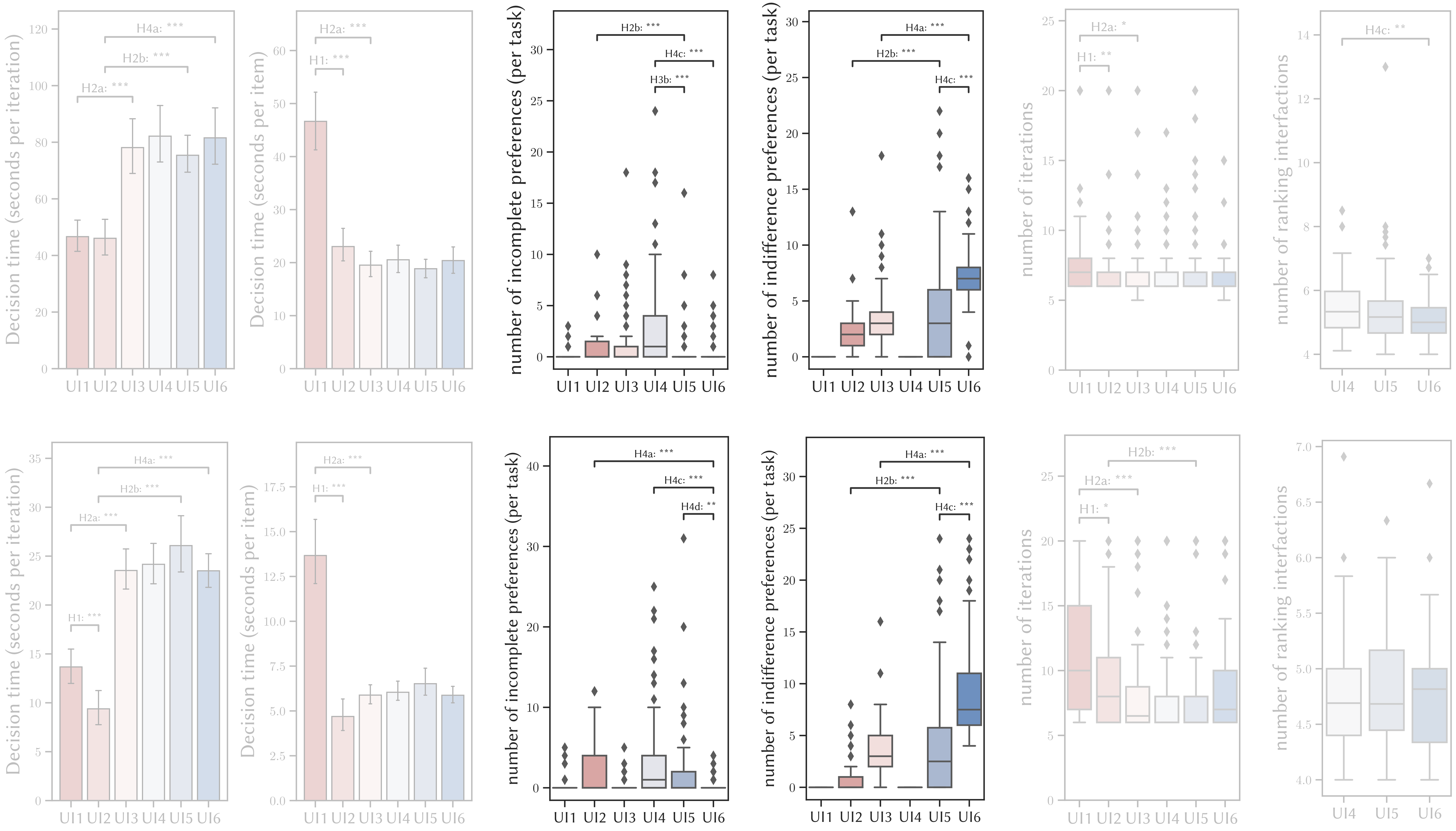
User Performance: Interaction Behavior

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

Participants express more incomplete and indifference judgements (UI2, UI3, UI4, UI5, UI6)

Text Summarization Task

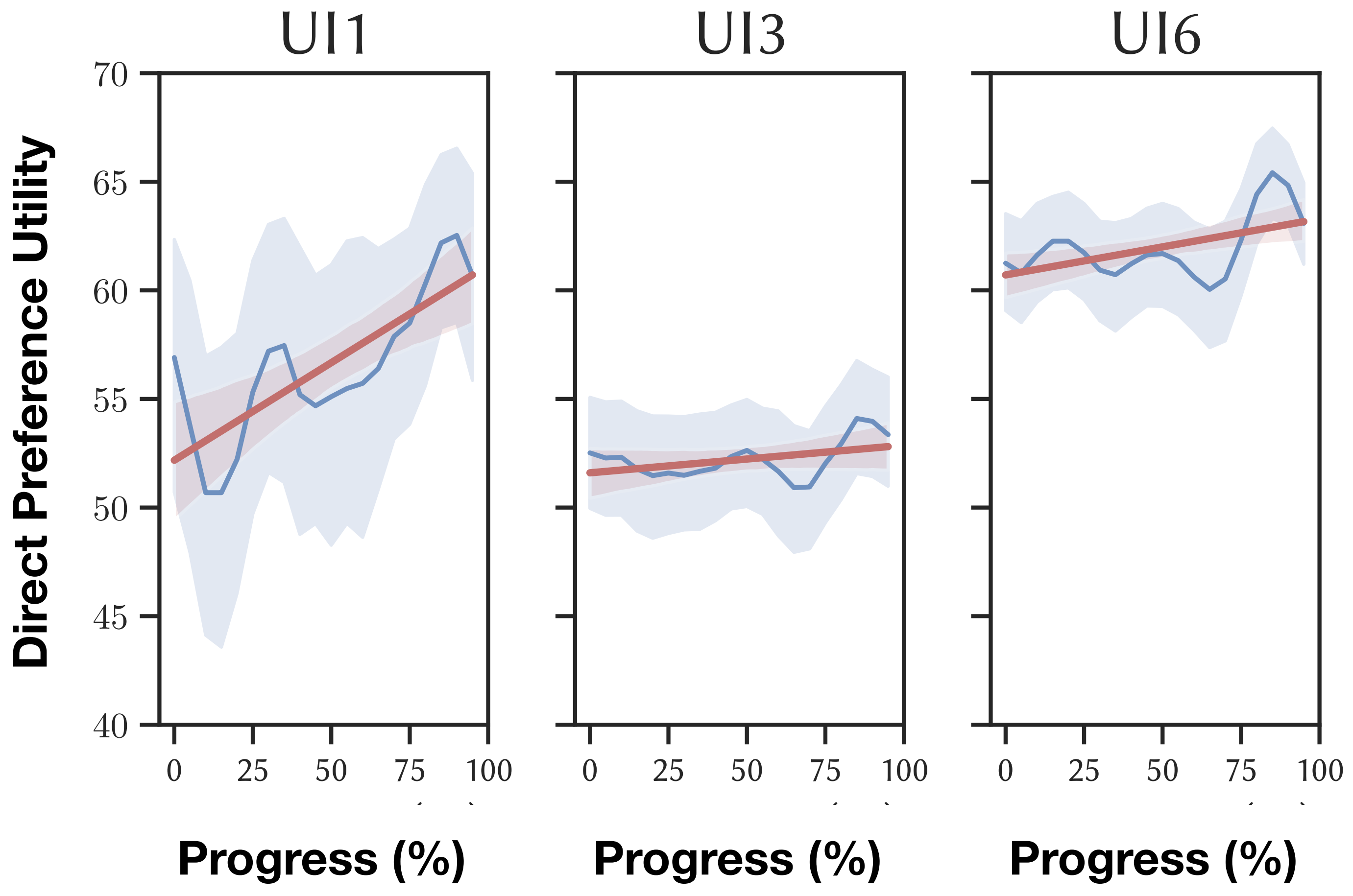
Photo Color Enhancement Task



Machine Performance: Utility (Direct and Latent)

UI6 significantly outperforms other UIs in terms of direct preference utility

Text Summarization Task



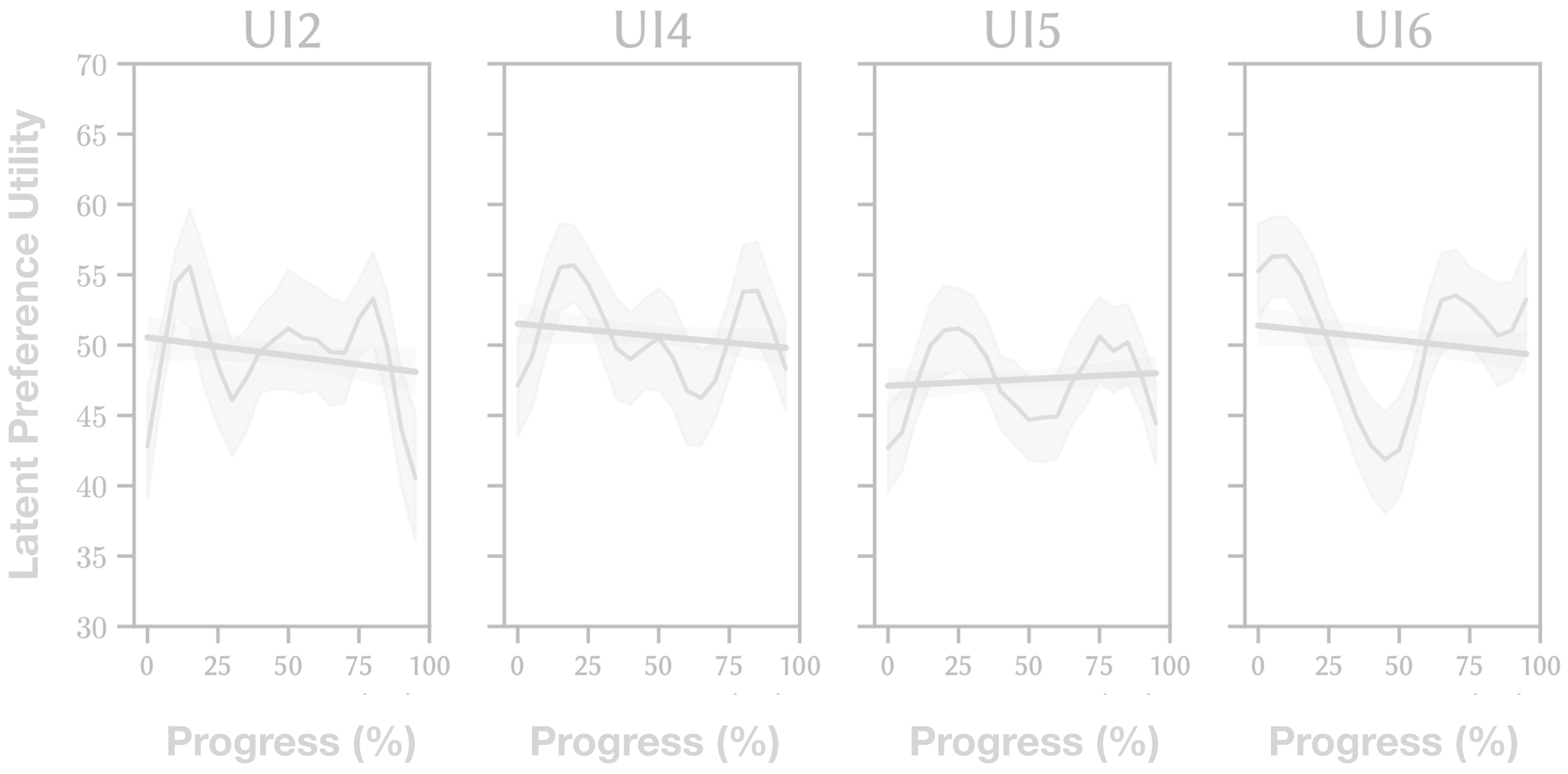
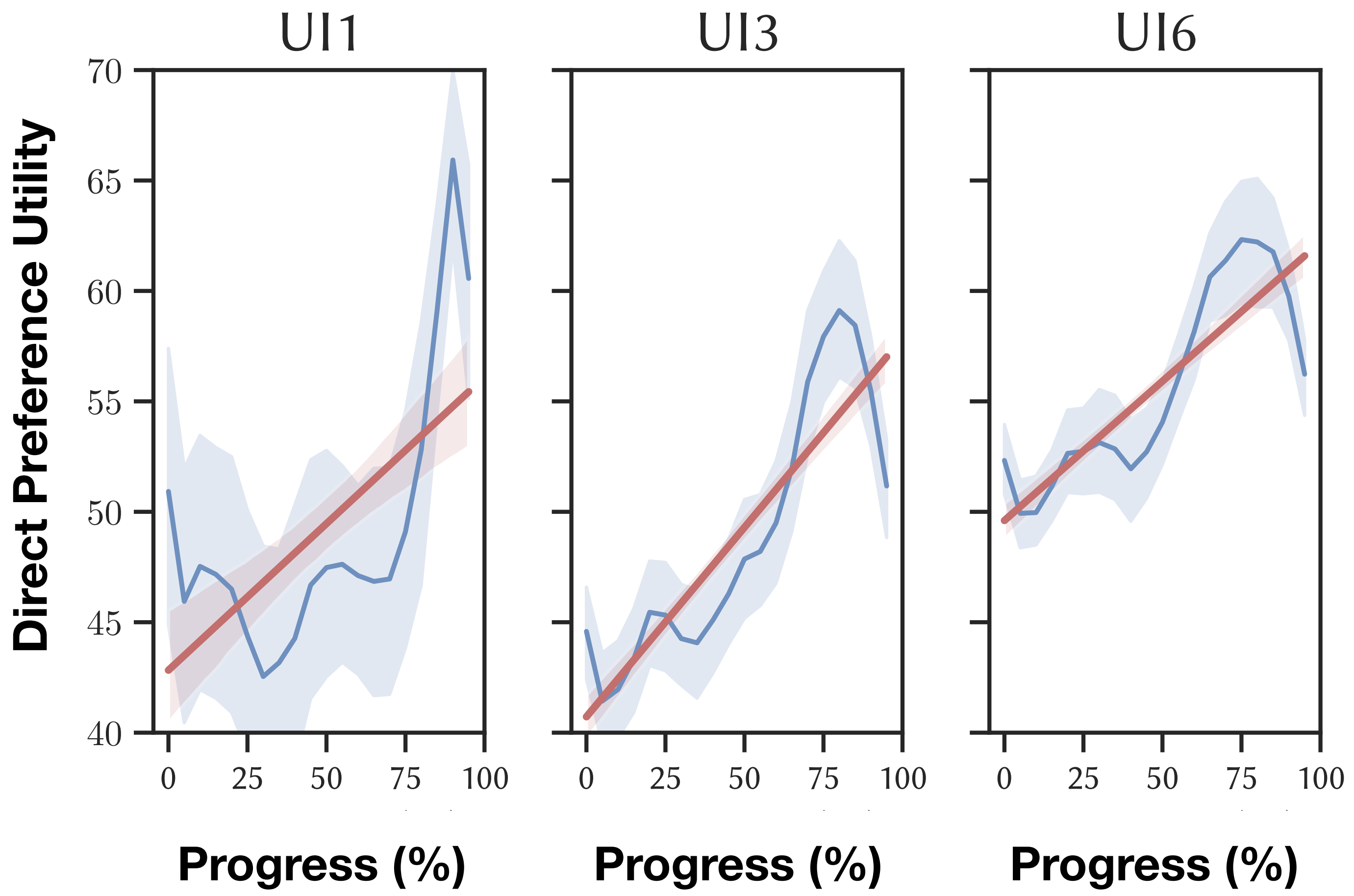
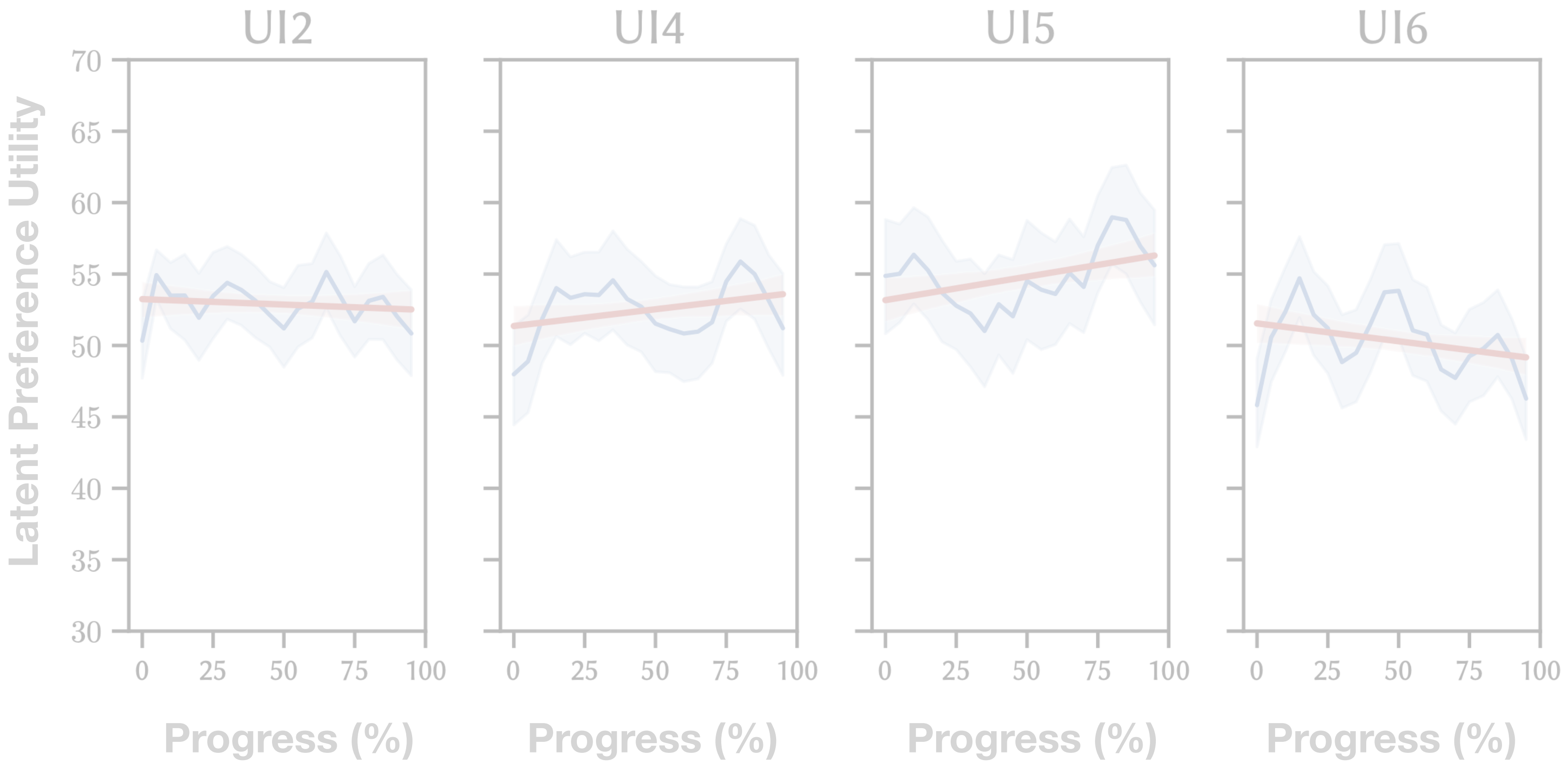


Photo Color Enhancement Task



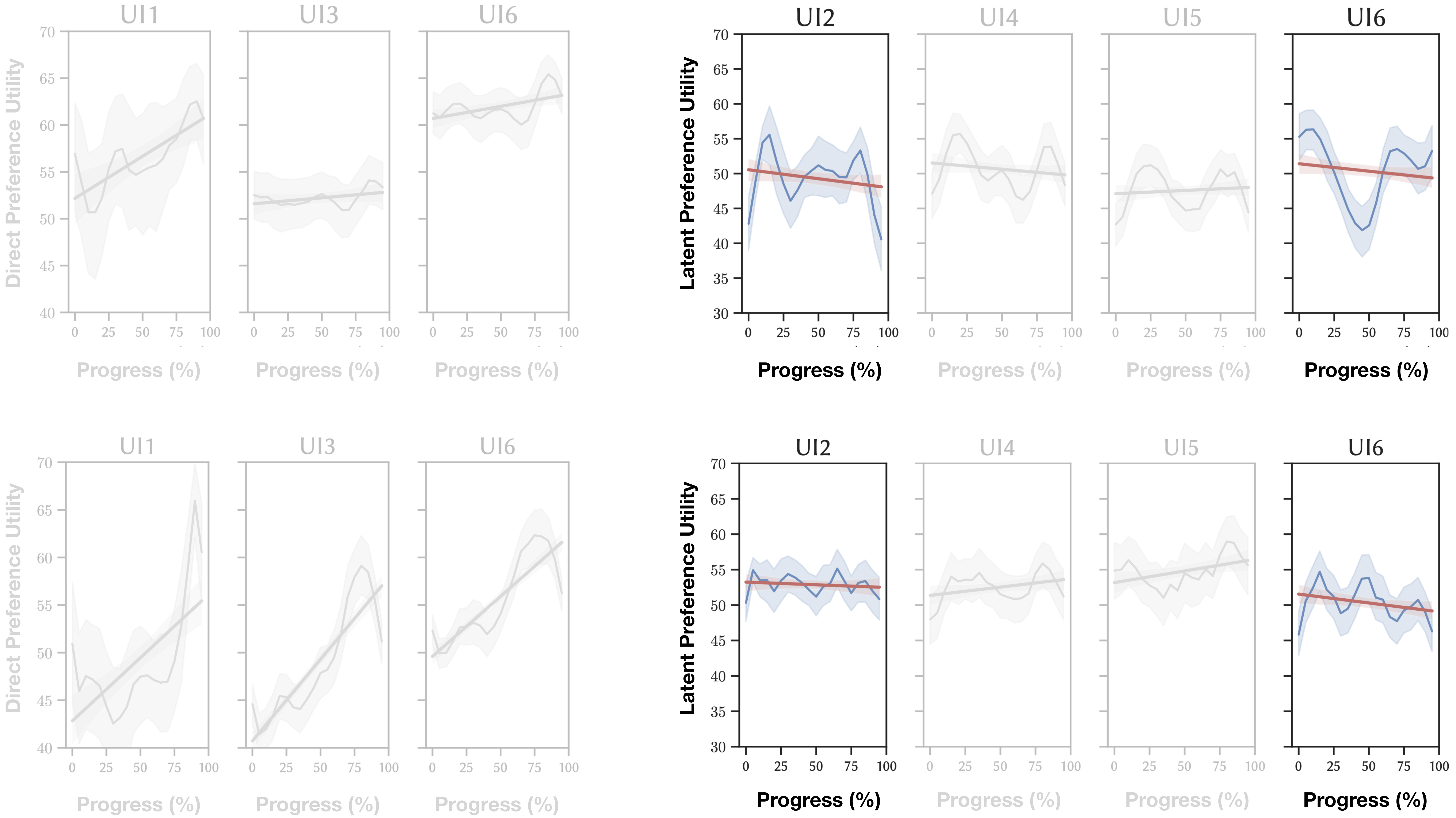


Machine Performance: Utility (Direct and Latent)

UI6 performed similarly compared to UI2 in terms of latent preference utility inferred by optimizer

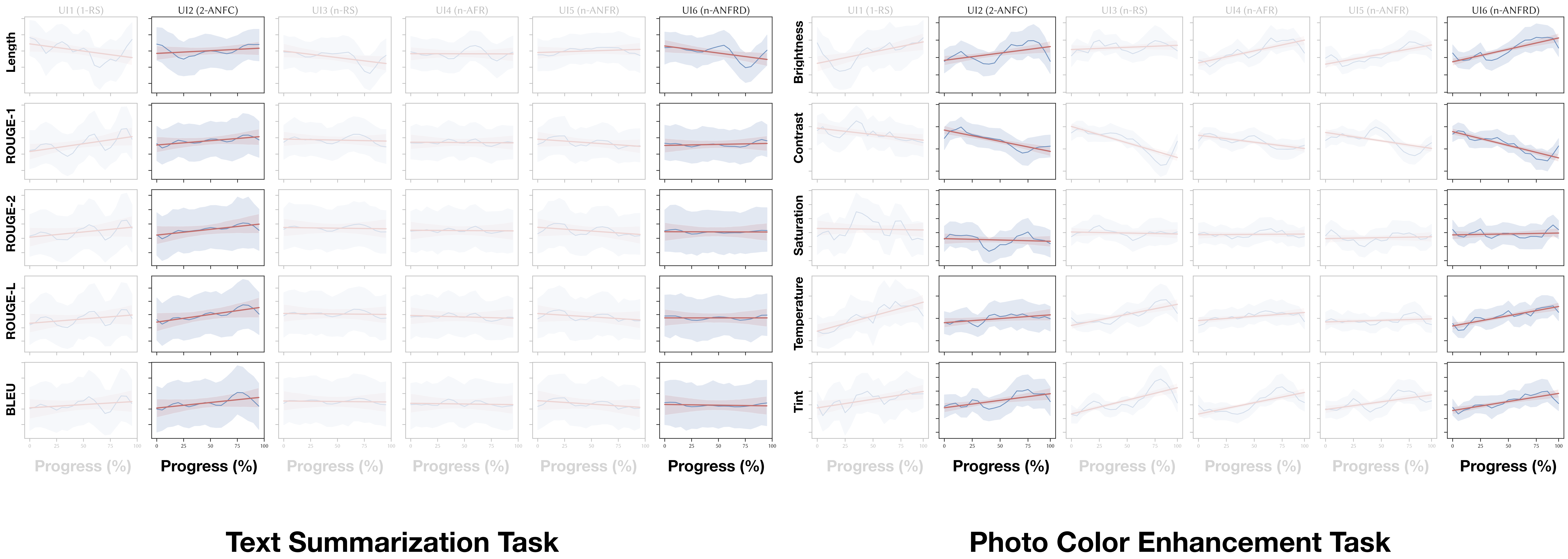
Text Summarization Task

Photo Color Enhancement Task



Machine Performance: Outcome (Objective) Quality

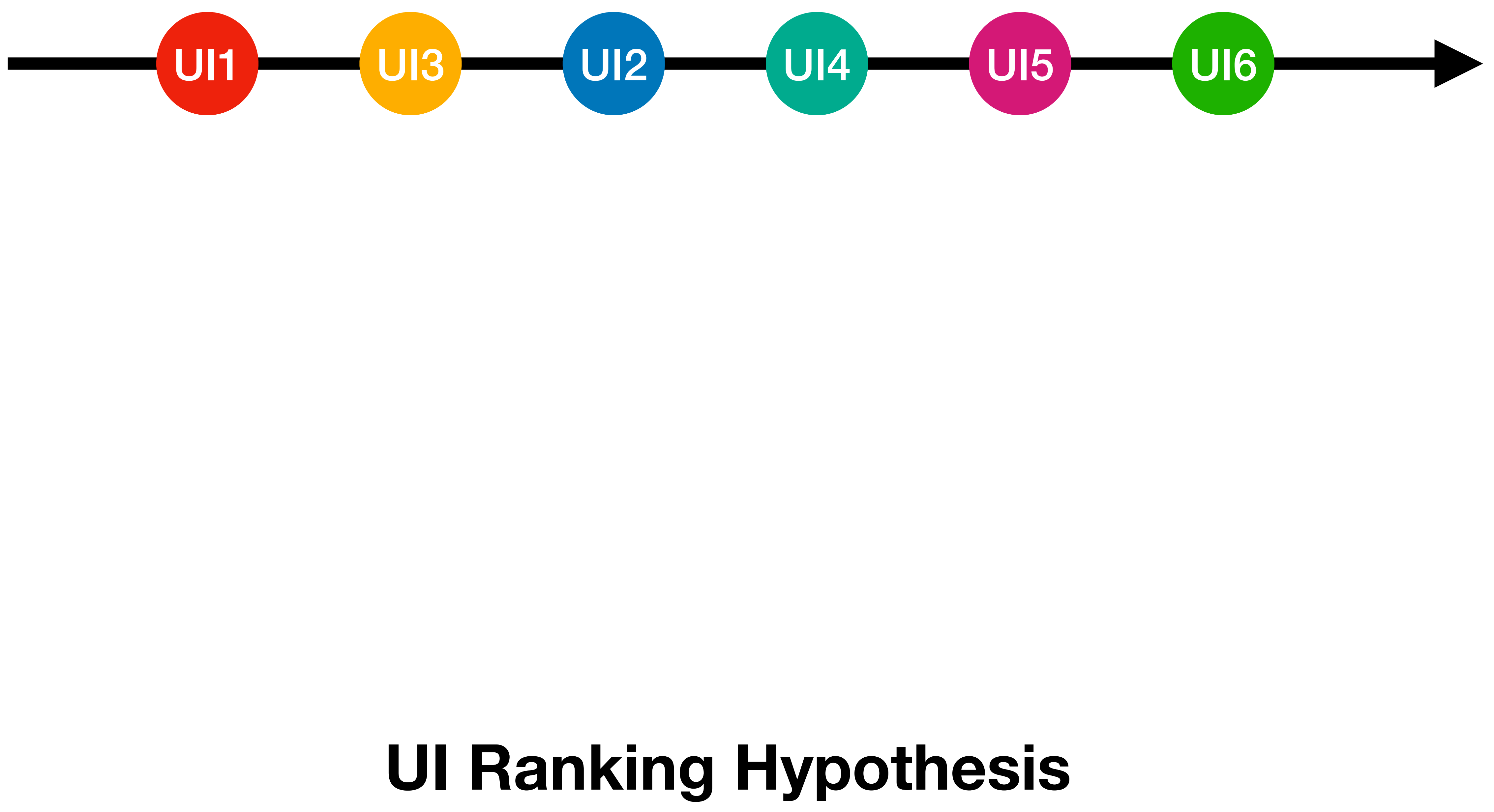
UI6 outperforms other UIs in terms of performance in various aspects

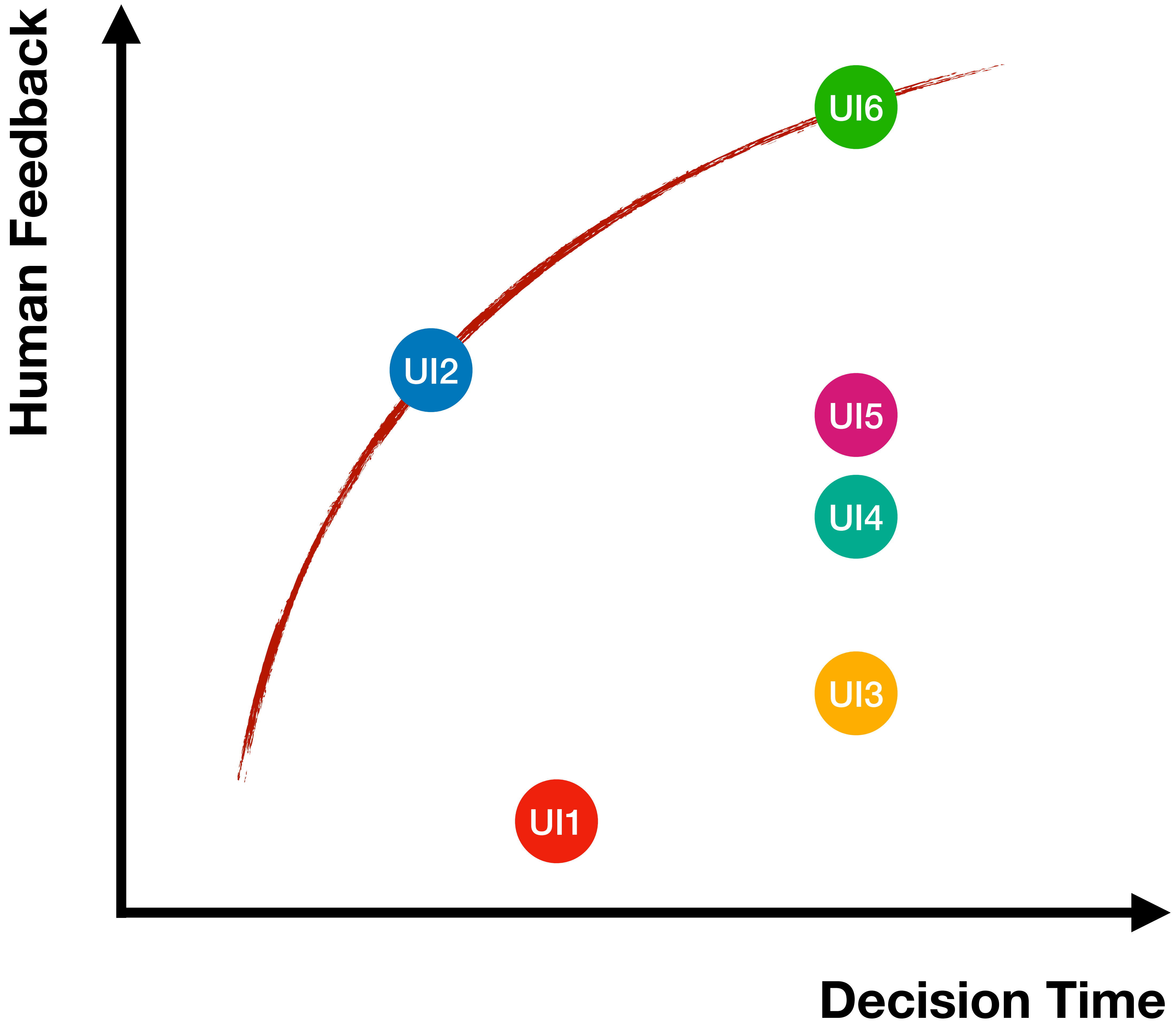


Opinion Measurement UI Performance

UI Ranking Hypothesis

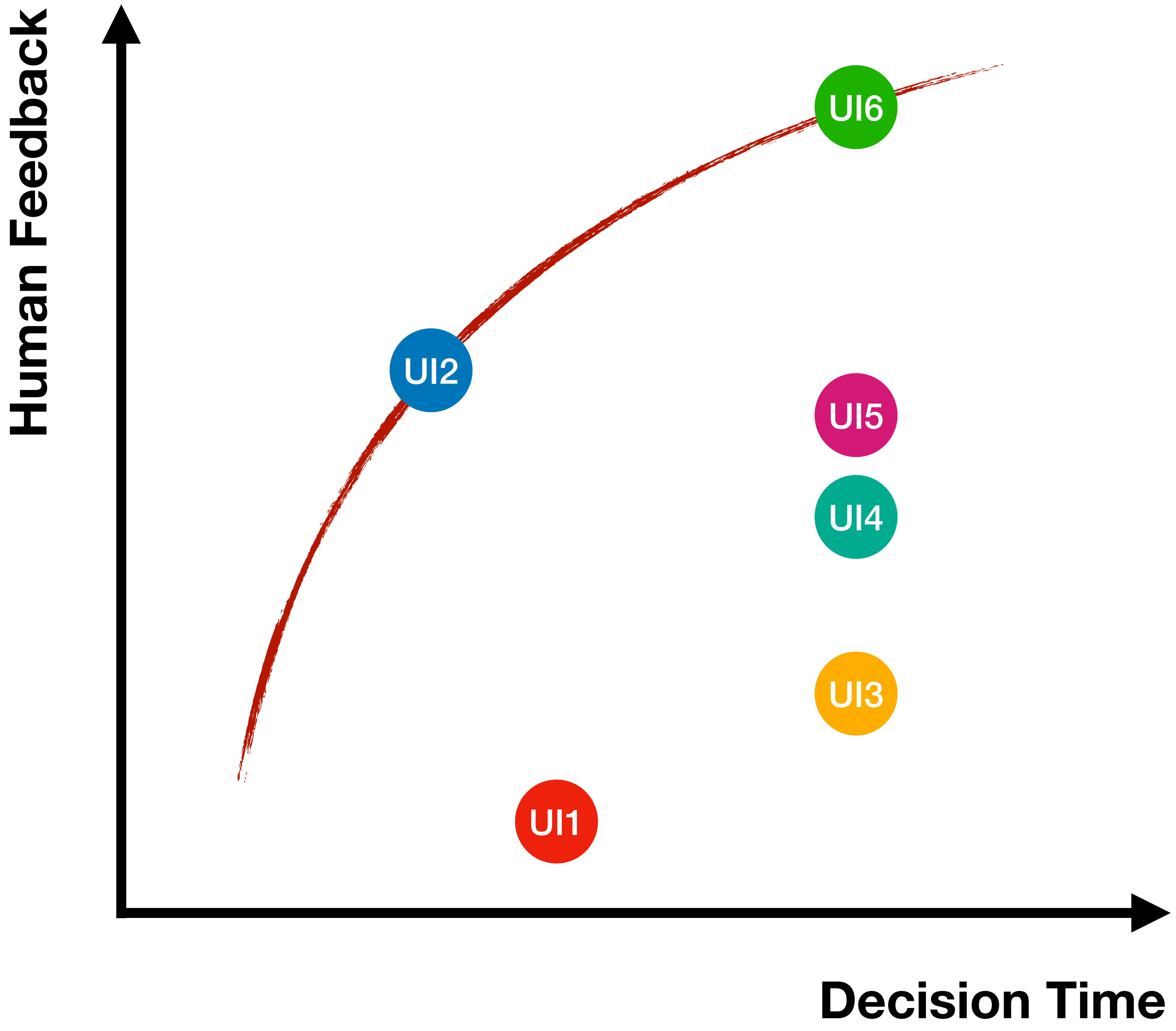
Opinion Measurement UI Performance





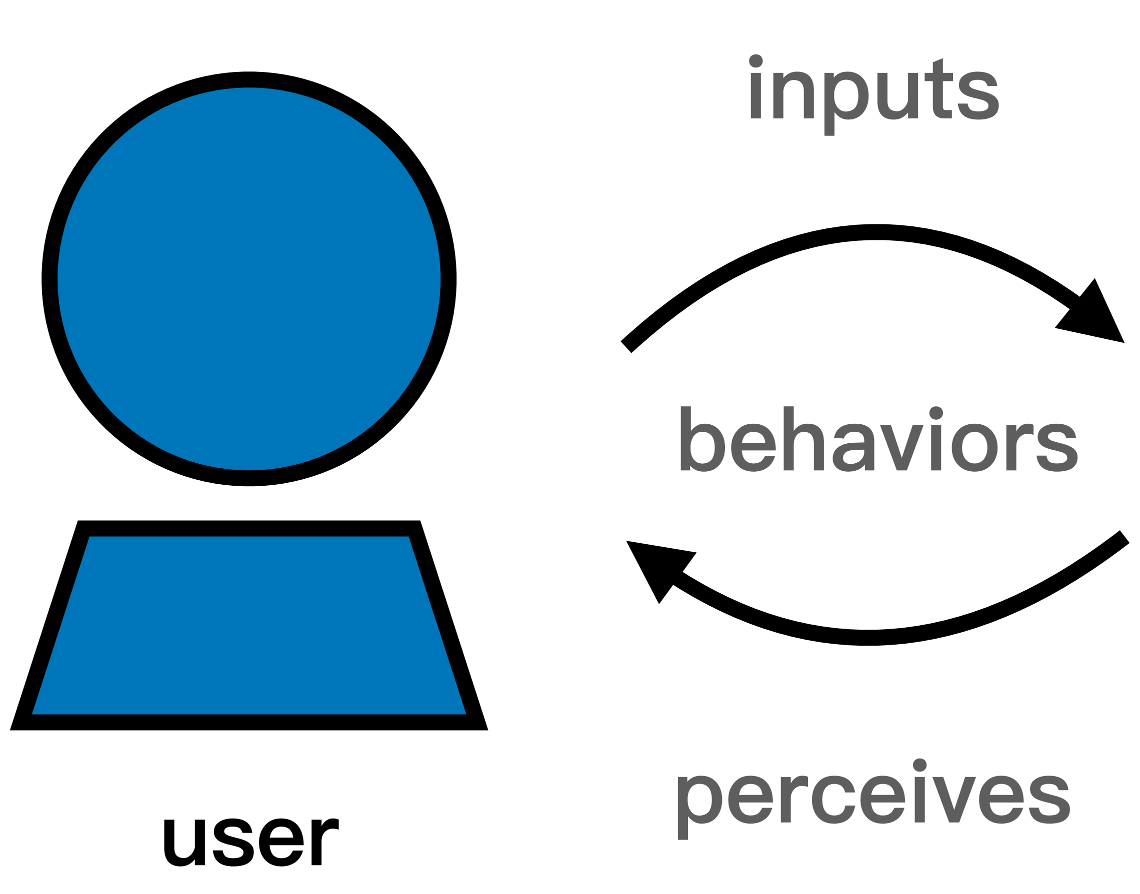
Optimal Opinion Measurement UIs for HITL

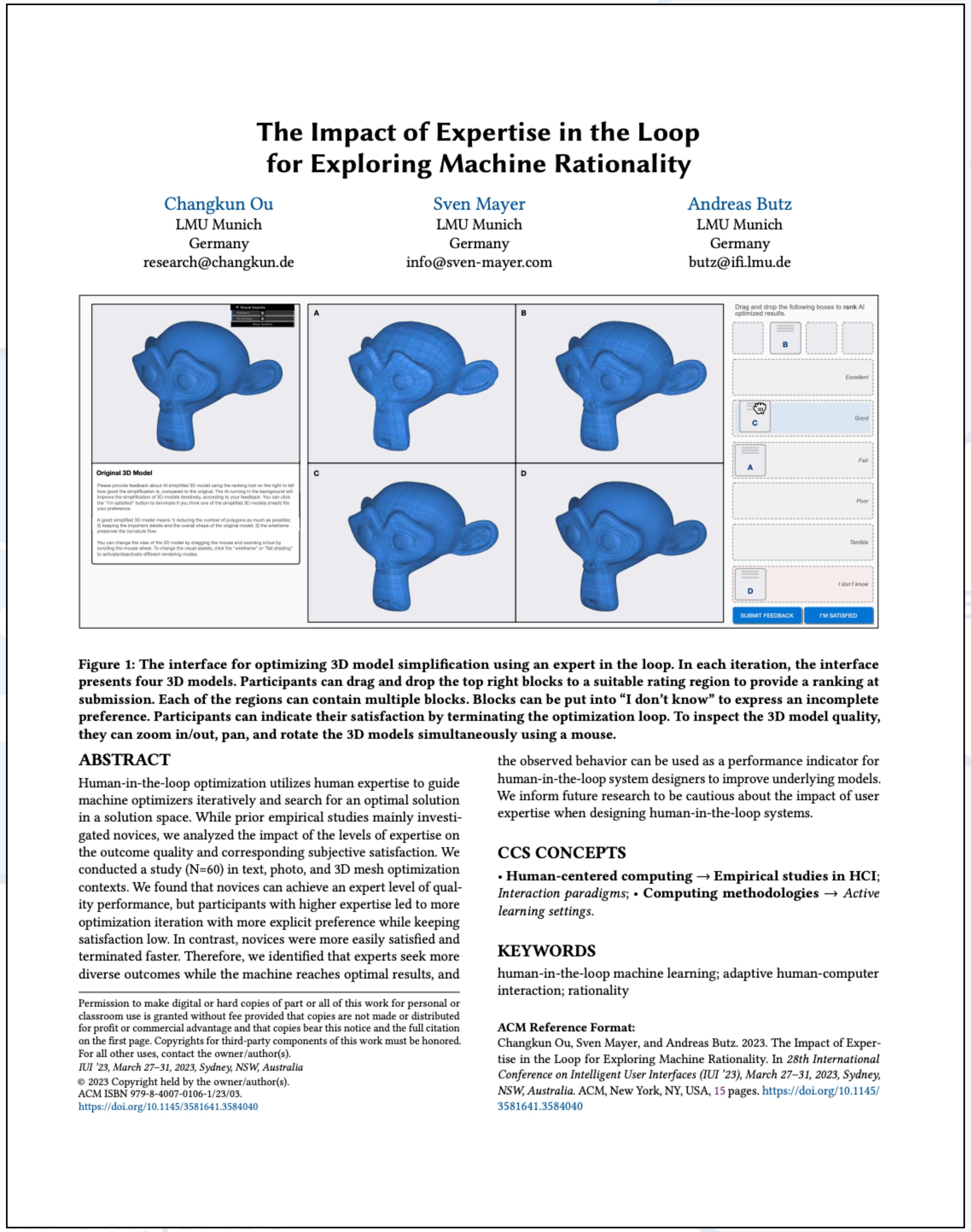
- Pointwise UI (UI1) should be avoided in general
- Pairwise UIs (e.g., UI2) favor fast decisions but are limited by consistent ranking order assumptions
- Listwise UIs without specified ranking distance (e.g. UI3, UI4, UI5) do not have general advantages over pairwise approach, but permitting a ranking 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

Changkun Ou
LMU Munich
Germany
research@changkun.de

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

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

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

ABSTRACT

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

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

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

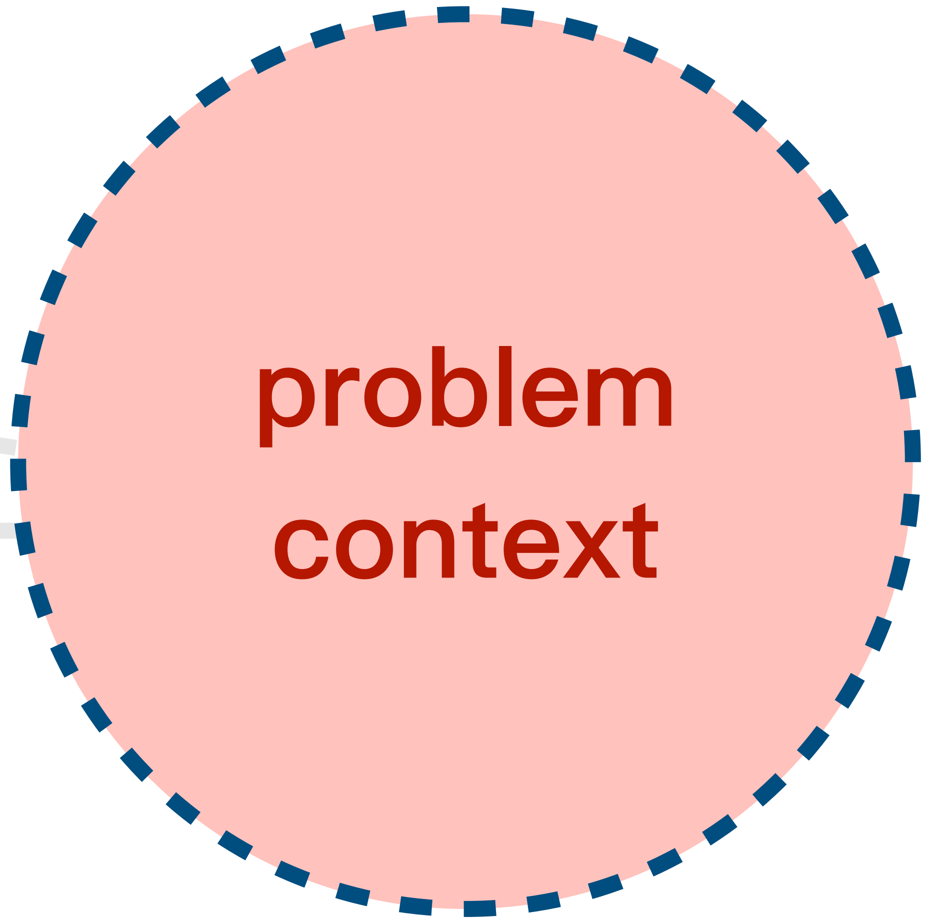
CCS CONCEPTS

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

KEYWORDS

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

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



Human Intelligence

Objective World

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:

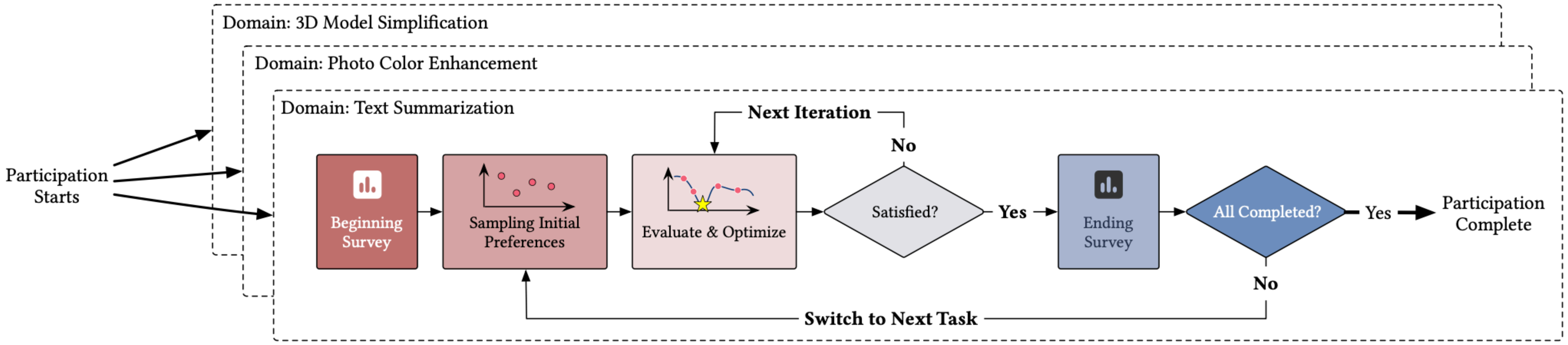
Fill beginning survey, providing feedback to the AI results; inspect improved results; loop until satisfaction, and fill ending survey

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)

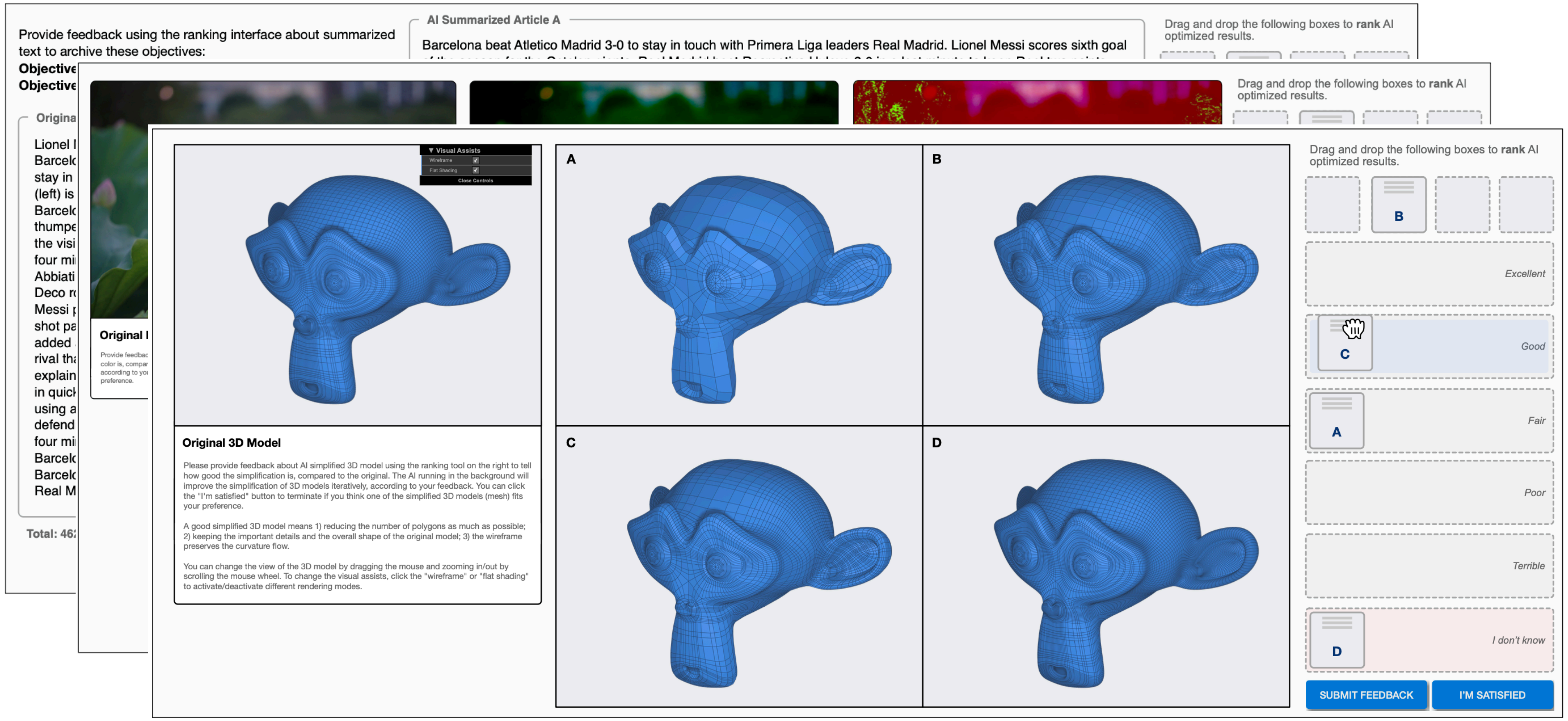


Apparatus

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

Photo color enhancer [Koyama et al. 2016, 2017, 2020]

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

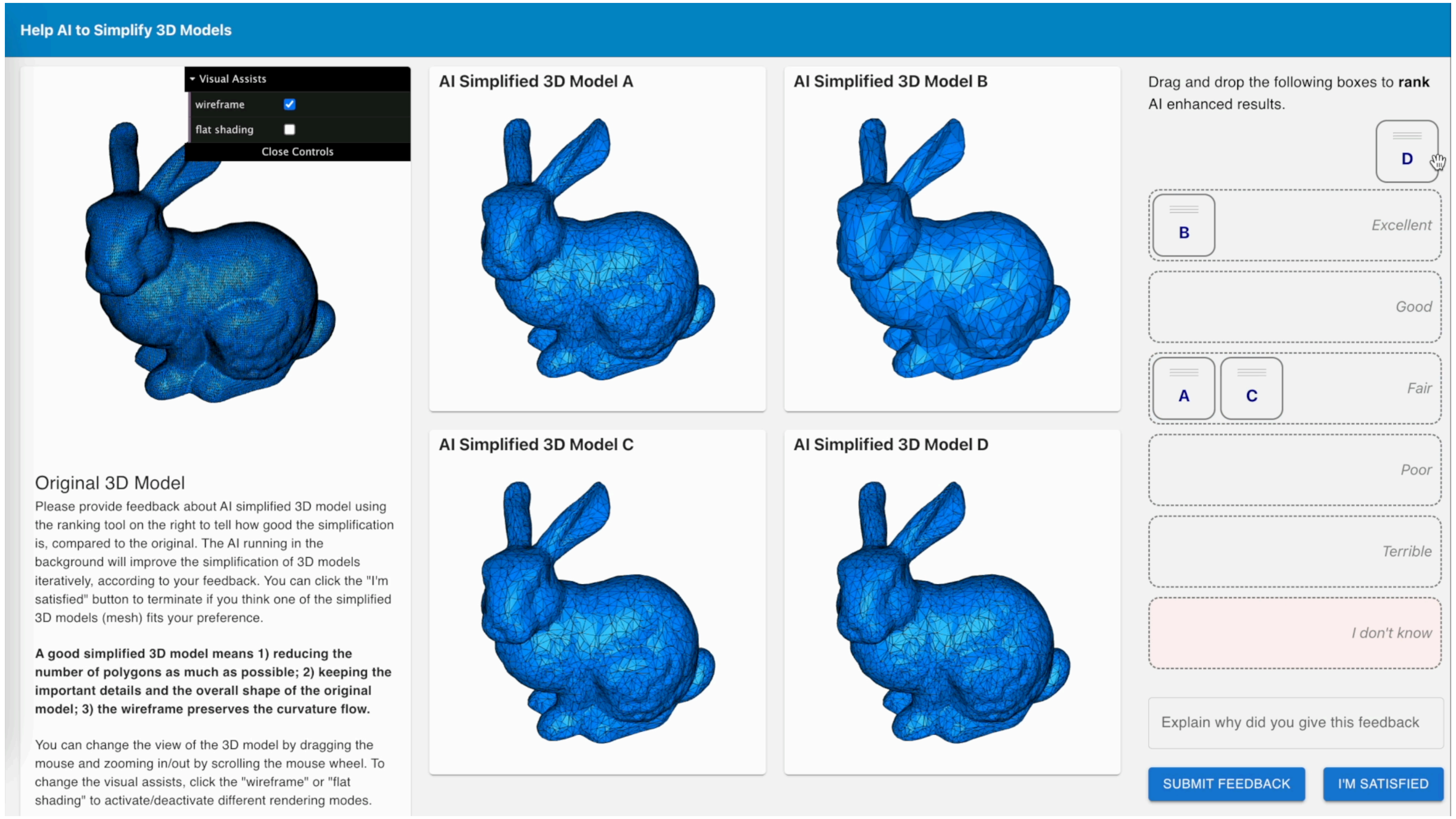


Apparatus: 3D Model Simplification

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*)

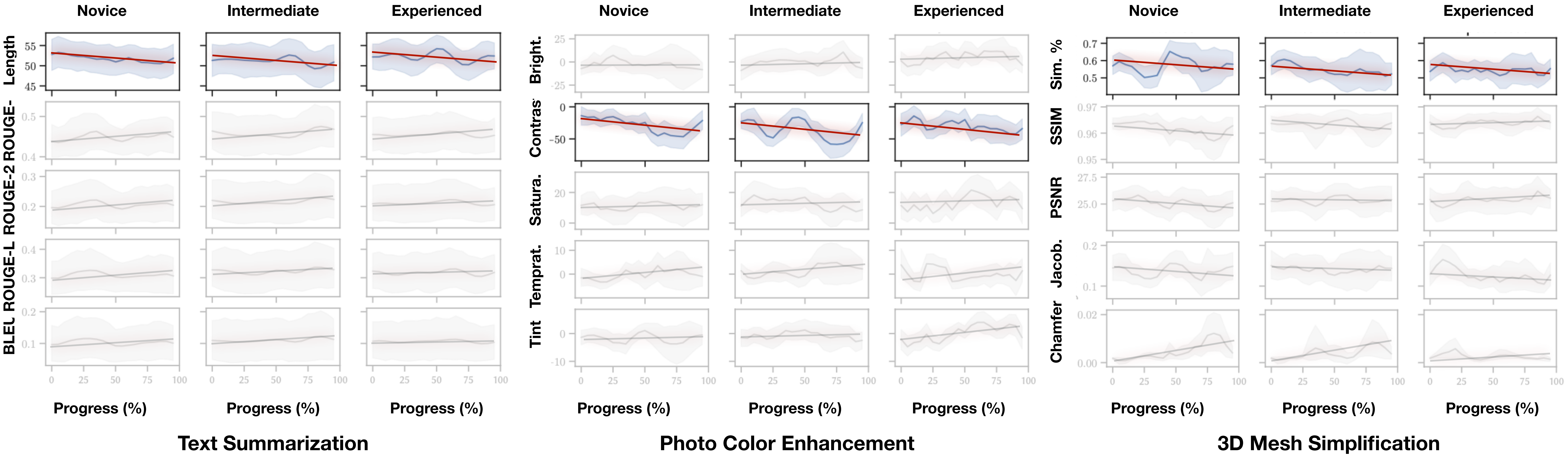
User performance

- 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: `perf ~ expertiseLevel * progress + (1|participant) + (1|task)`

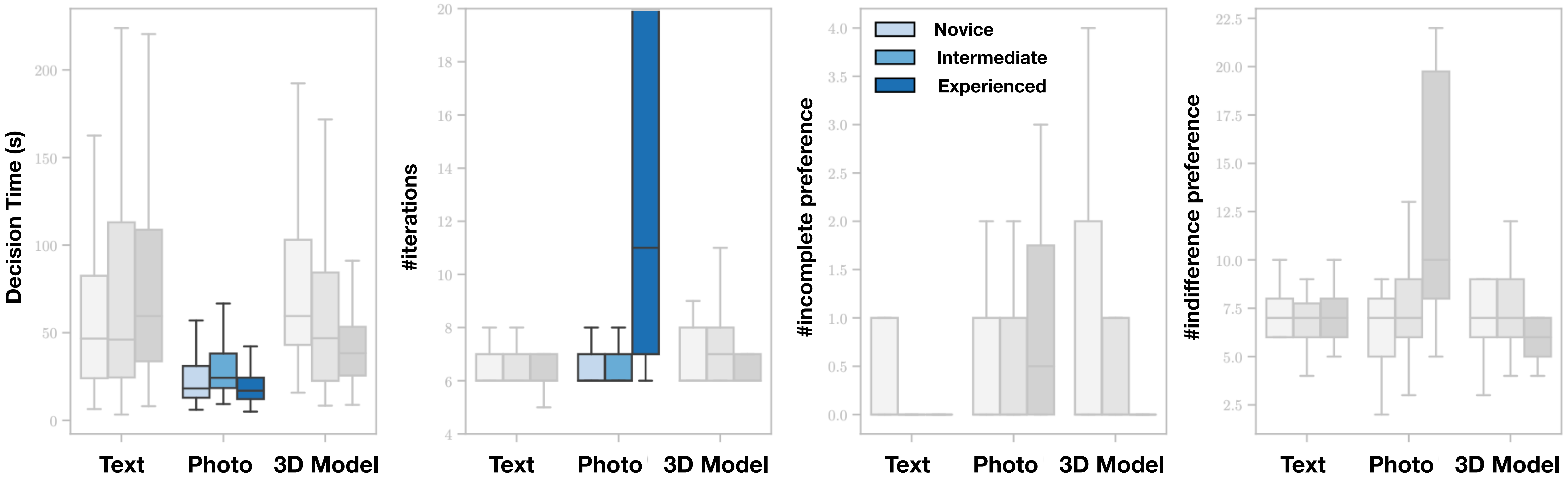
Machine Performance: Objective Outcome Quality

Novices and intermediates can reach expert level performance



User Performance: Interaction Behavior

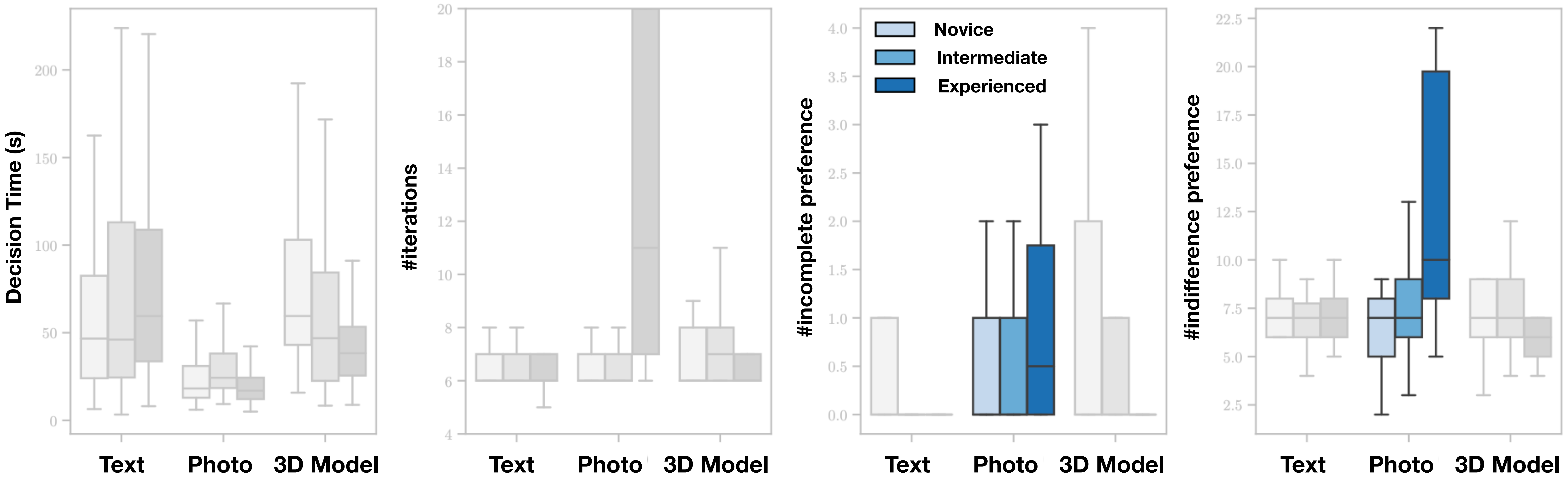
Experienced participants explore solutions more when feedback loop is more efficient



User Performance: Interaction Behavior

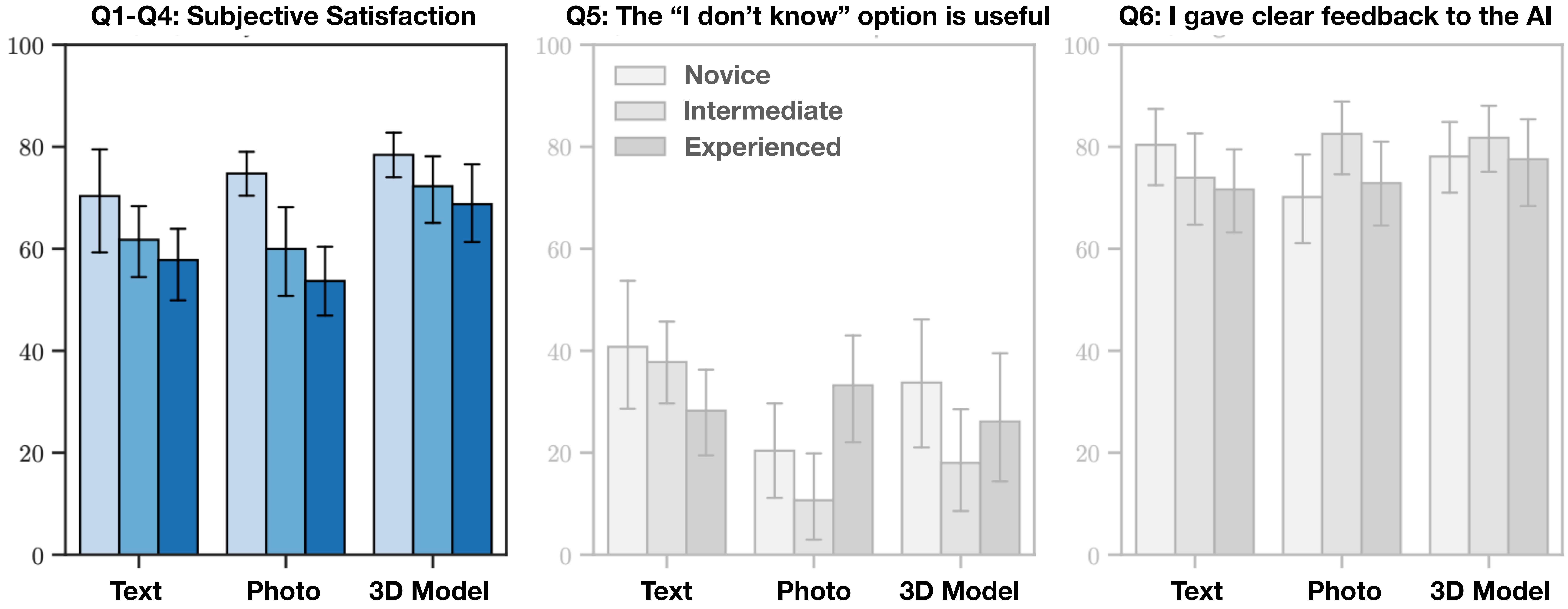
Experienced participants explore solutions more when feedback loop is more efficient

Experienced participants indicate a clearer preference by showing more frequently incomplete/indifferent preferences

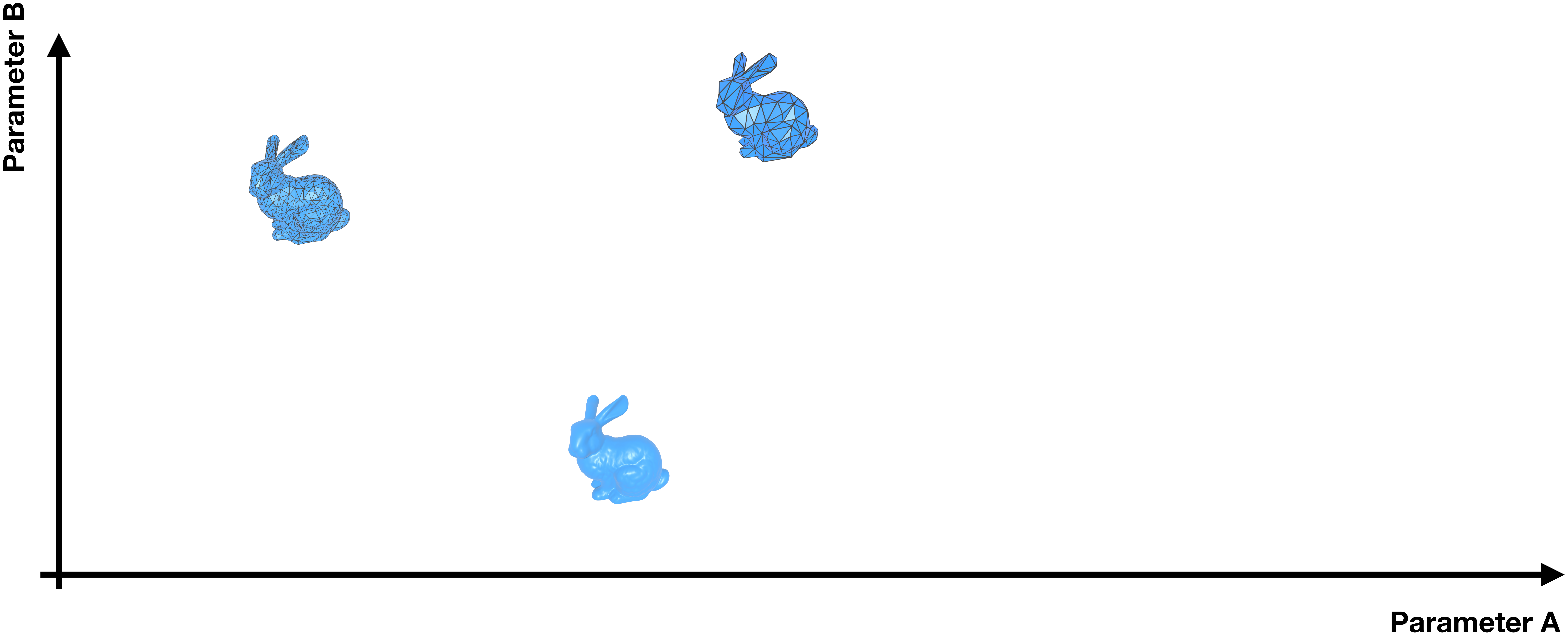


User Performance: Subjective Satisfaction

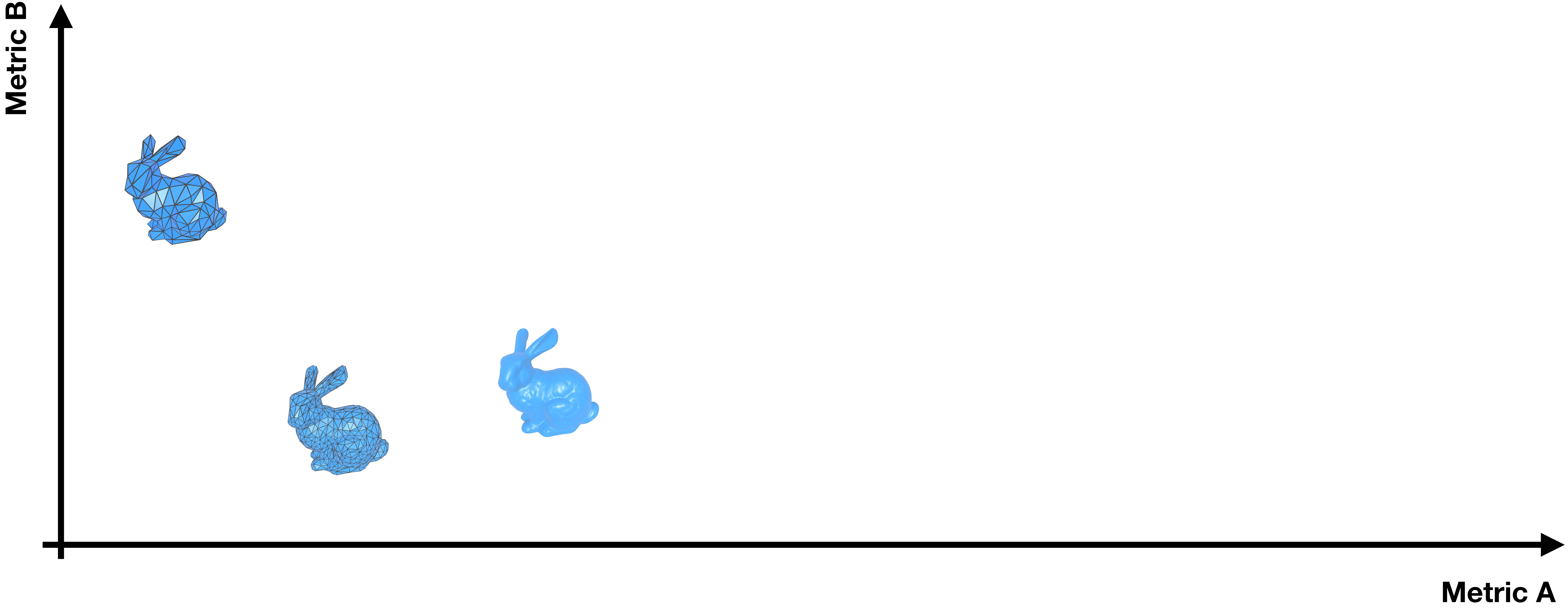
Instead, **novices** are significantly **more satisfied** than experienced ones



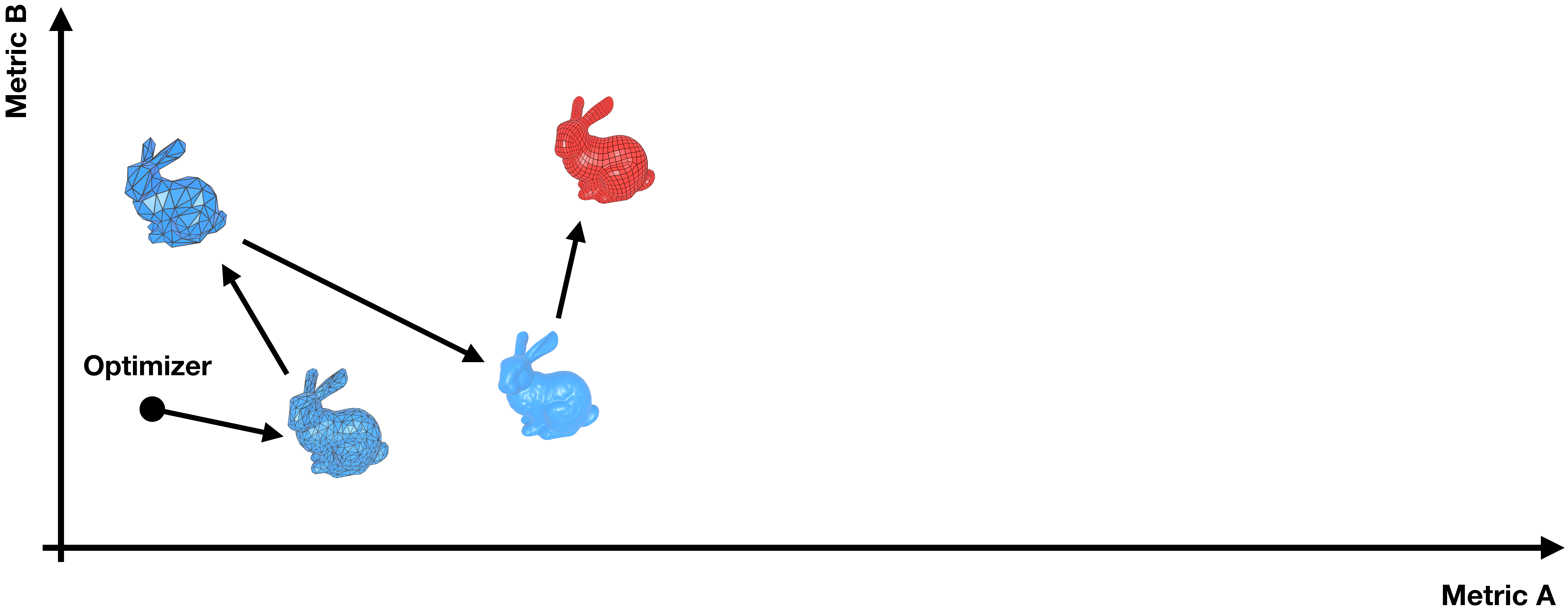
Solution Space



Performance Space



Performance Space

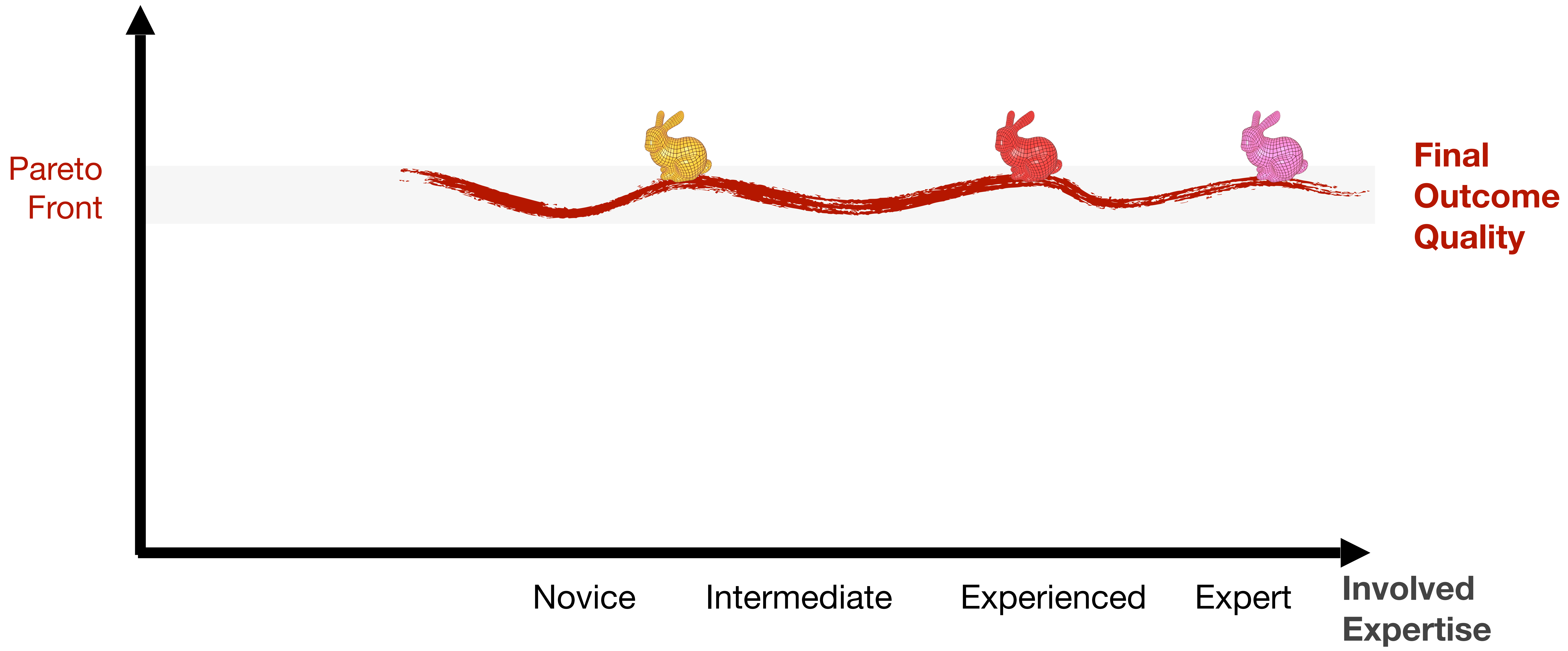


Performance Space

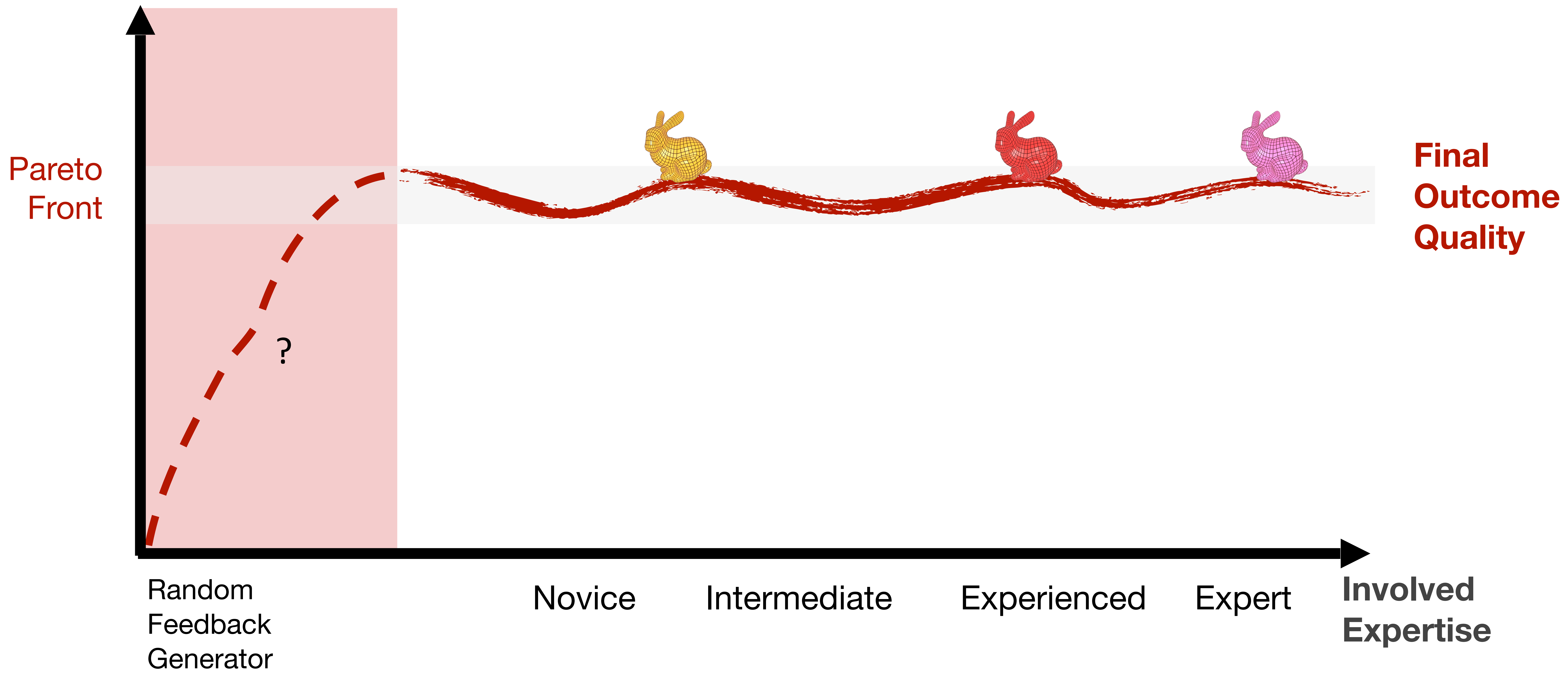
Performance Space

Performance Space

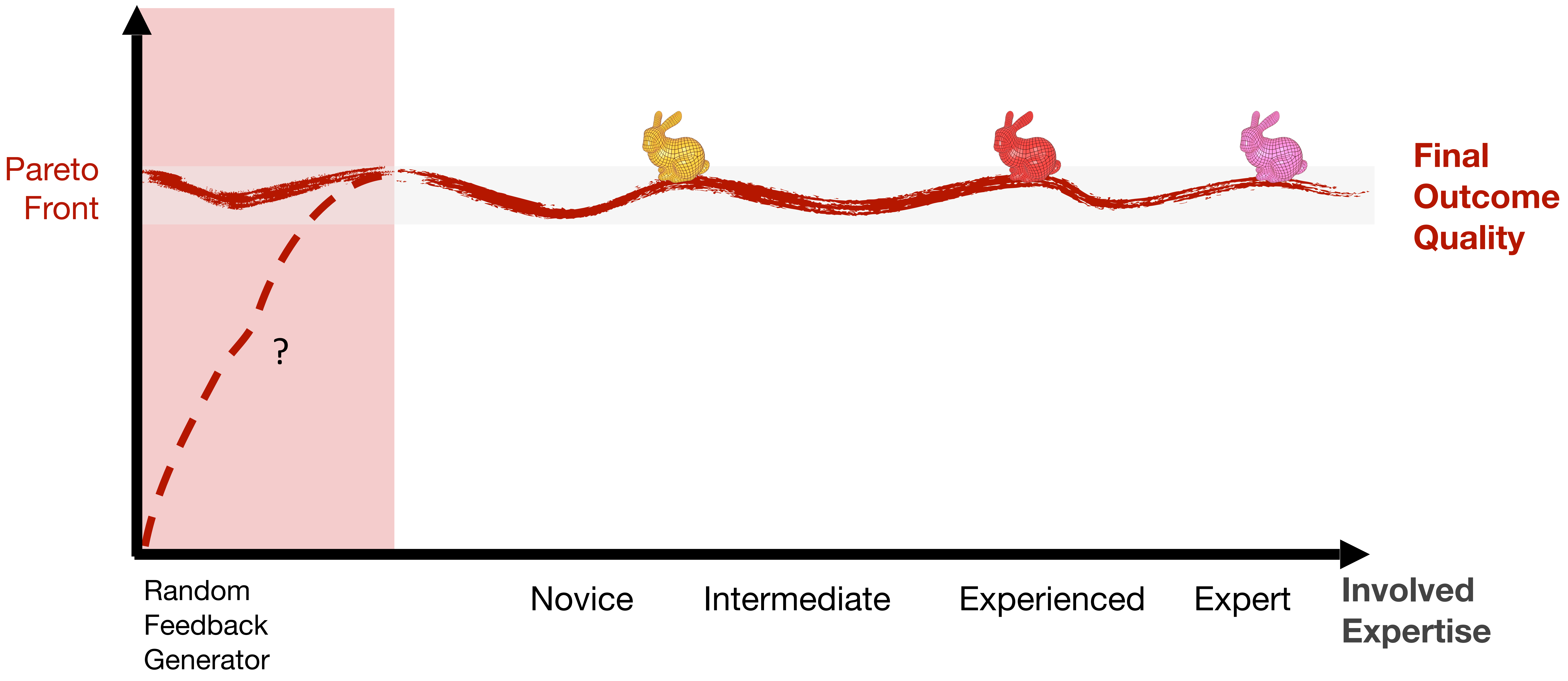
Implications on the Impact of Expertise in HITL



Implications on the Impact of Expertise in HITL



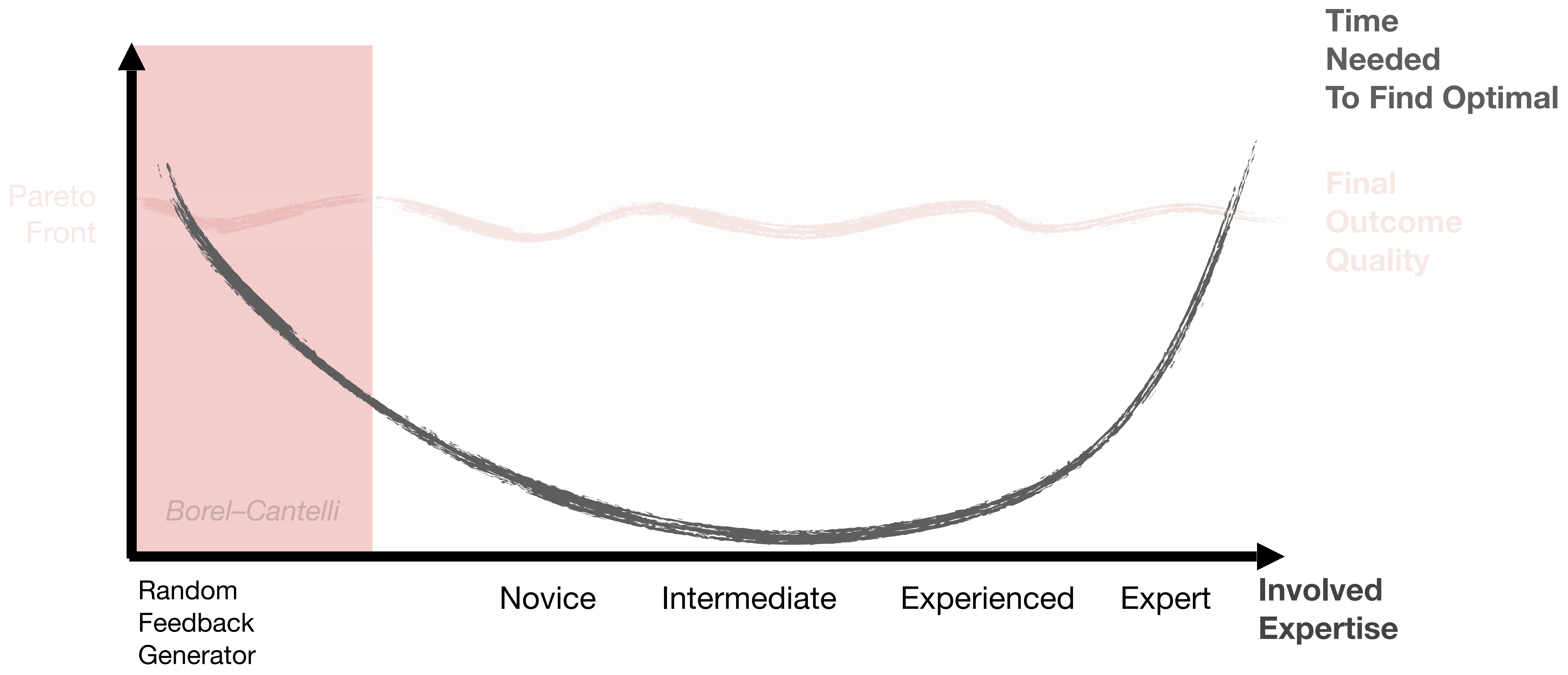
Implications on the Impact of Expertise in HITL



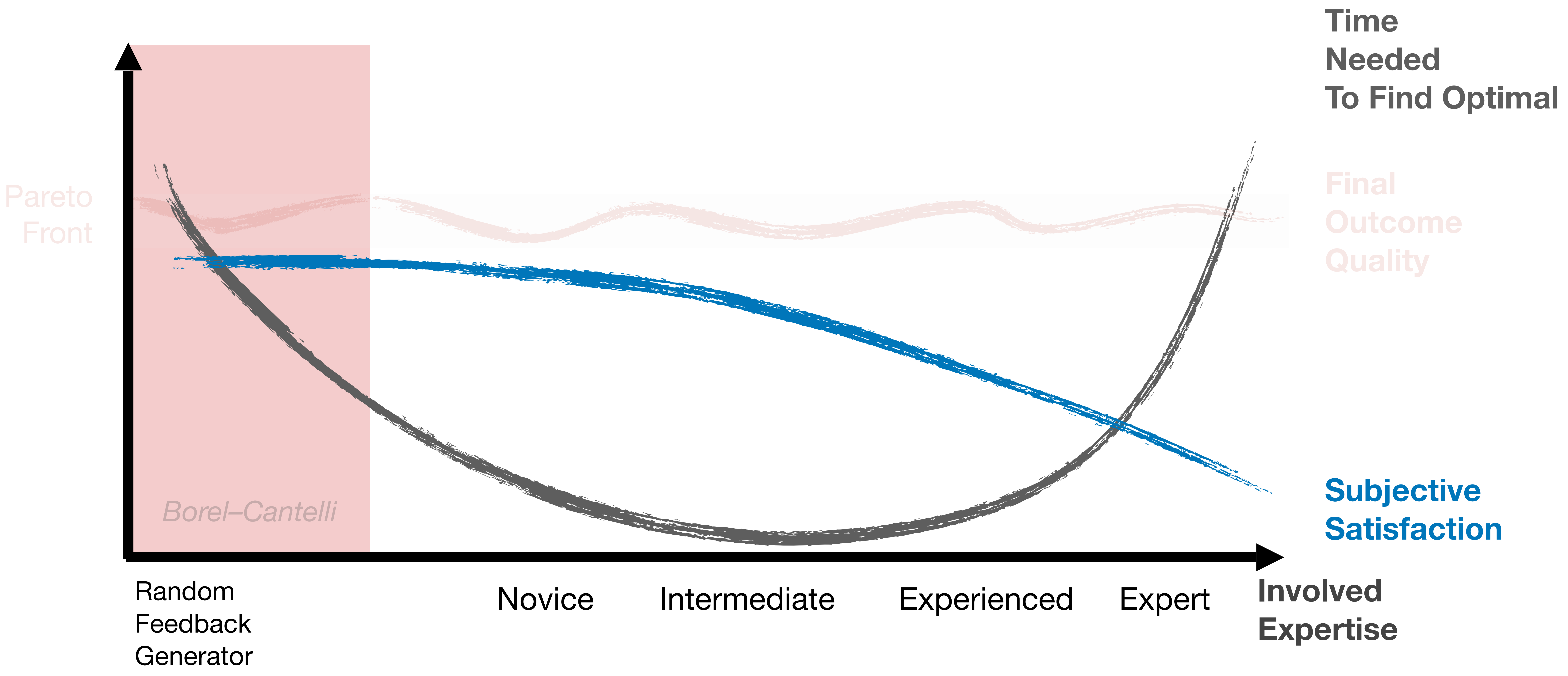
The Borel–Cantelli lemma [Borel 1909] [Cantelli 1917]
With infinite amount of events, the probability of observing any meaningful result is 1.0

$$\sum_{n=1}^{\infty} P(E_n) = +\infty \Rightarrow P(\{E_n \text{i.o.}\}) = 1$$

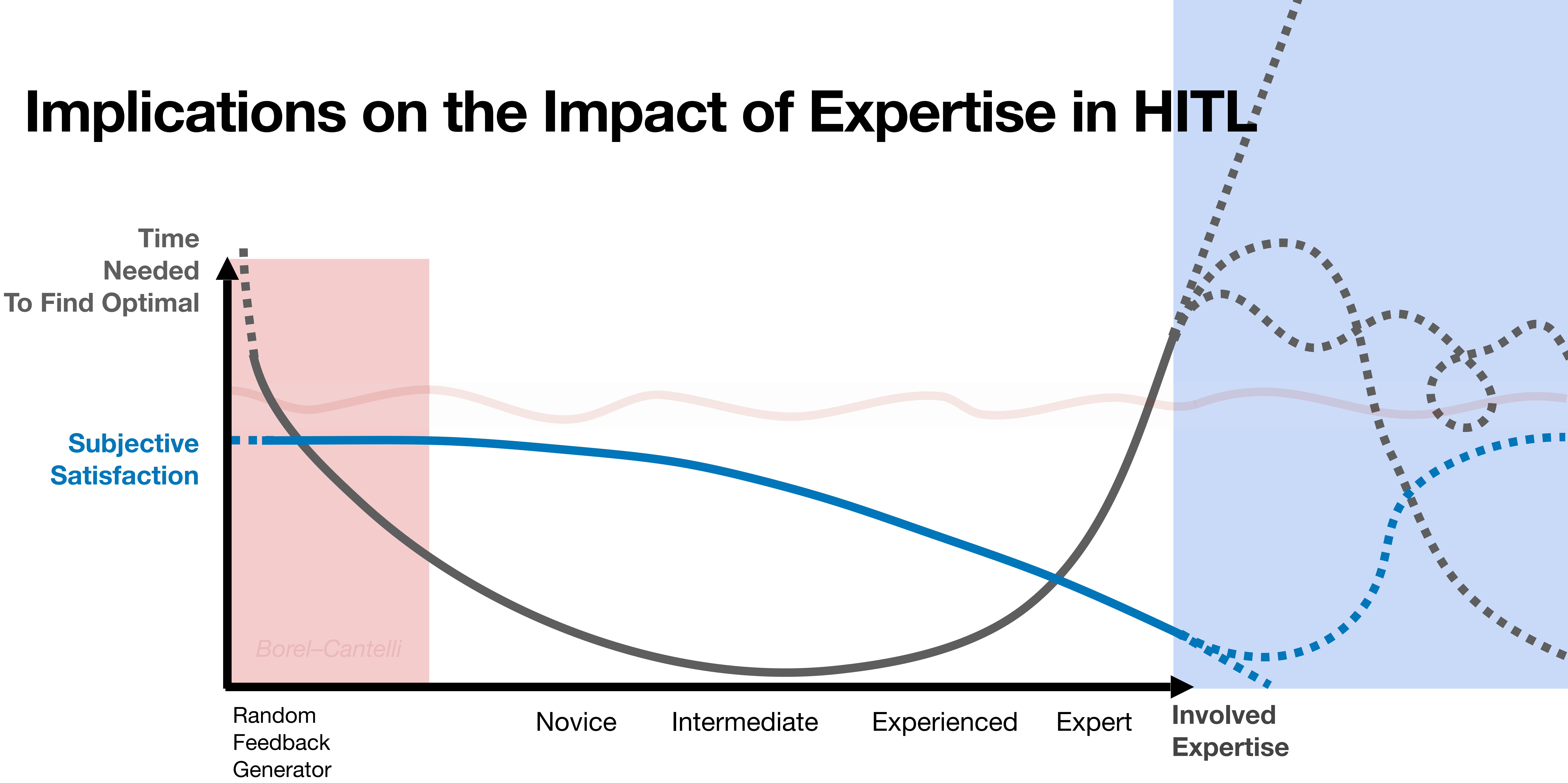
Implications on the Impact of Expertise in HITL



Implications on the Impact of Expertise in HITL



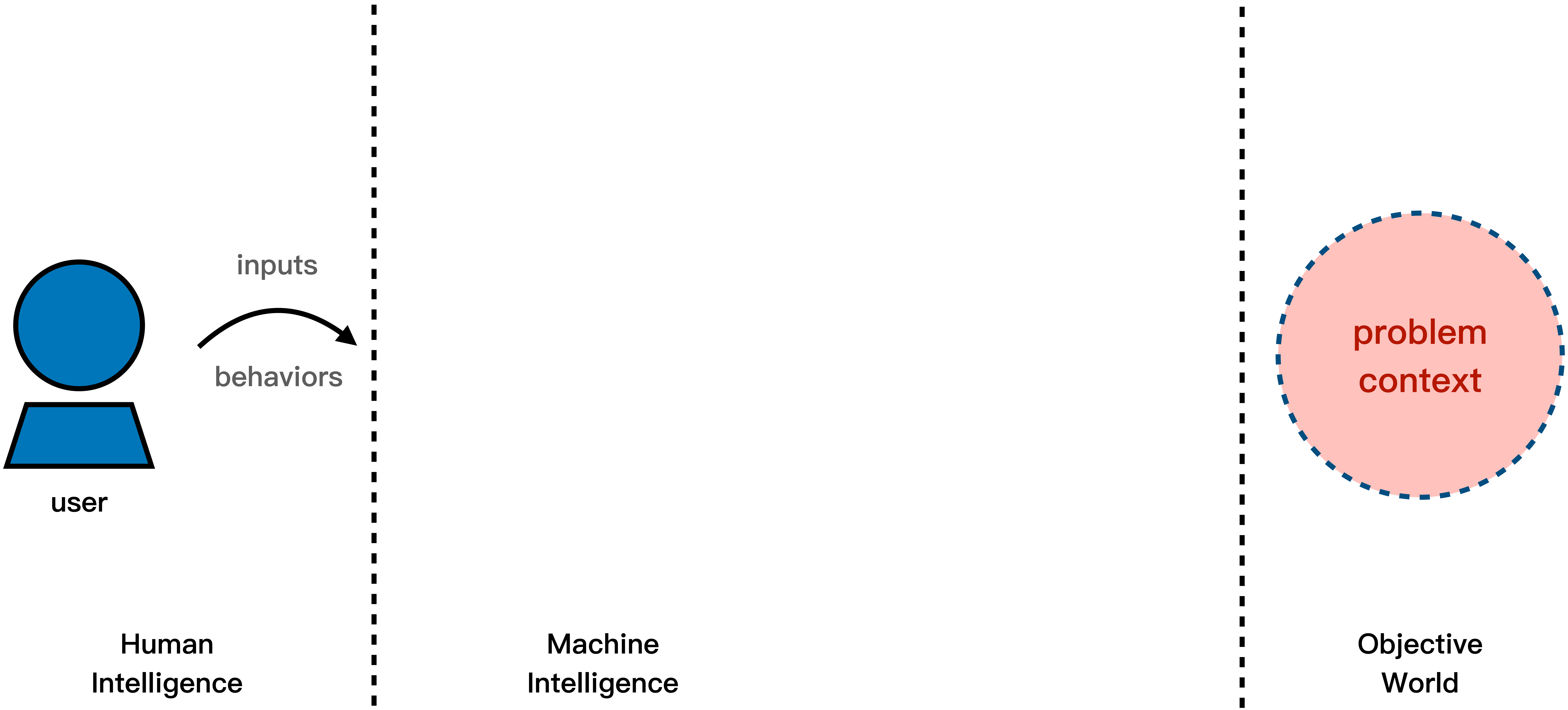
Implications on the Impact of Expertise in HITL



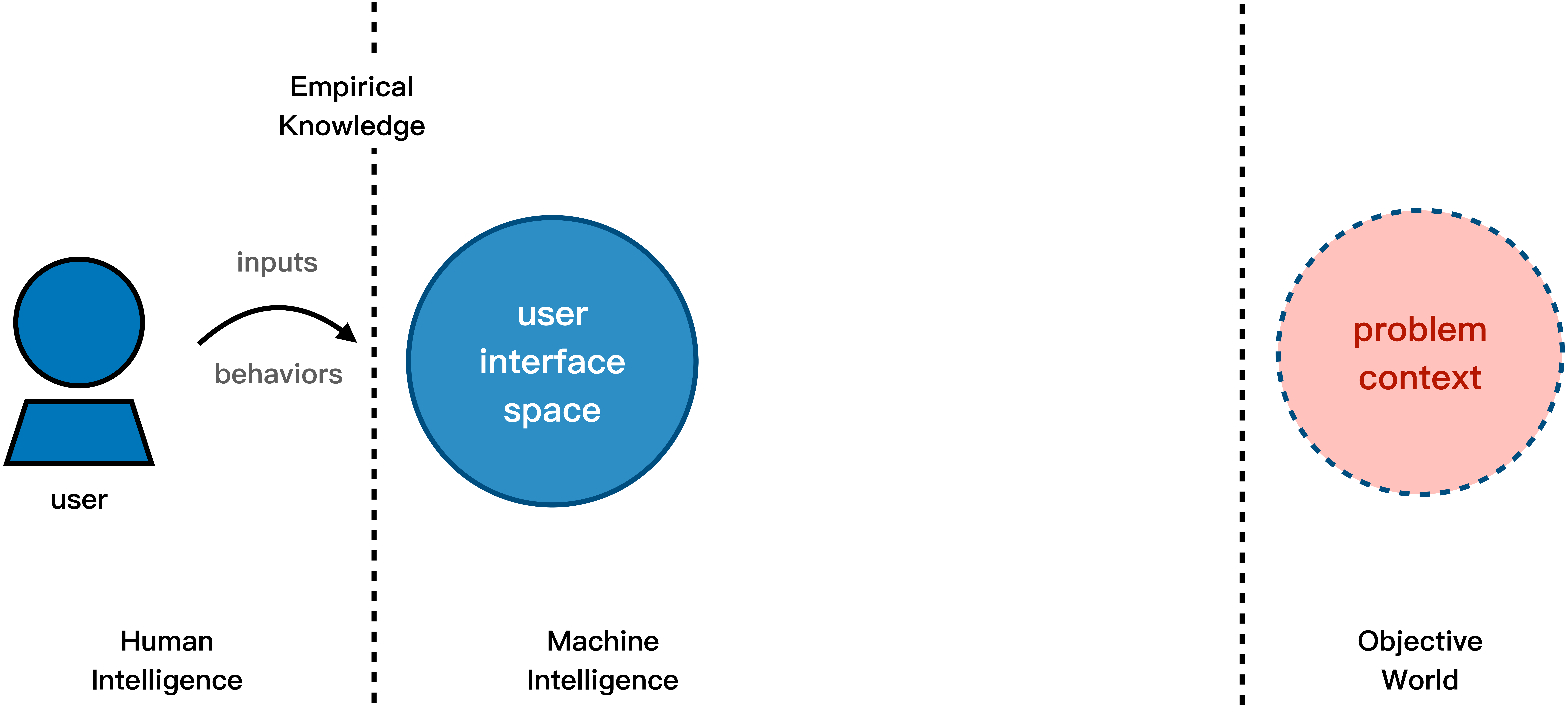
Reflections

Building Blocks in HITL Optimization Systems

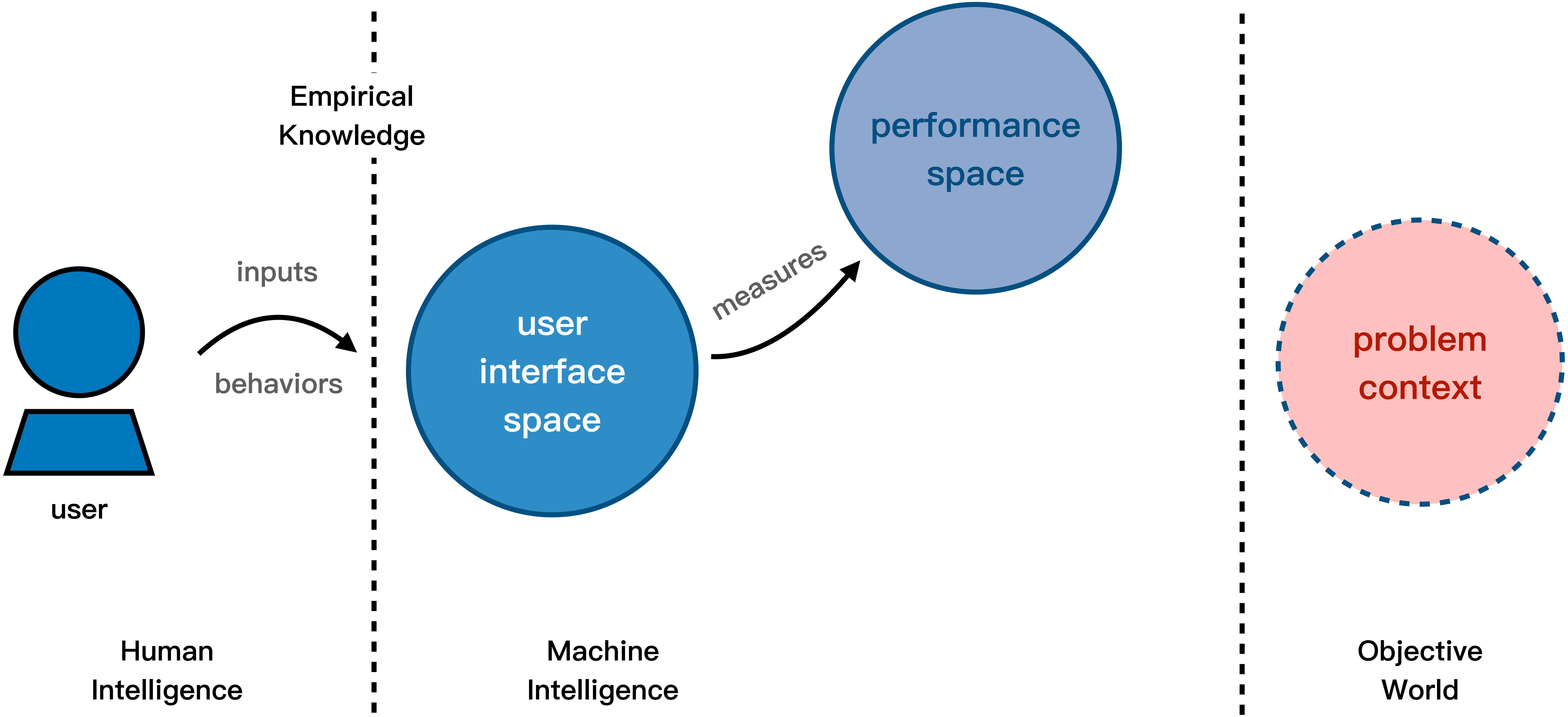
Building Blocks in HITL Optimization Systems



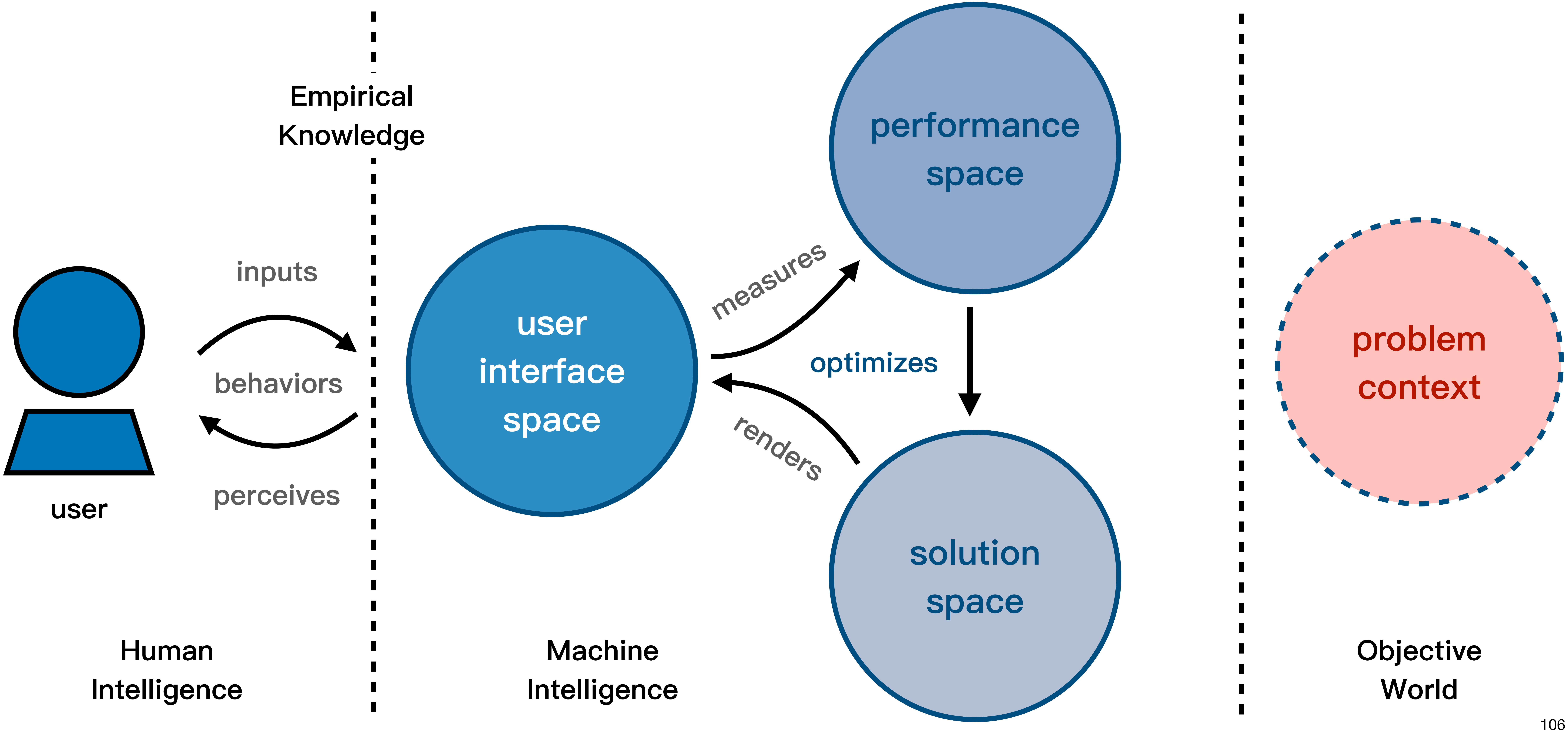
Building Blocks in HITL Optimization Systems



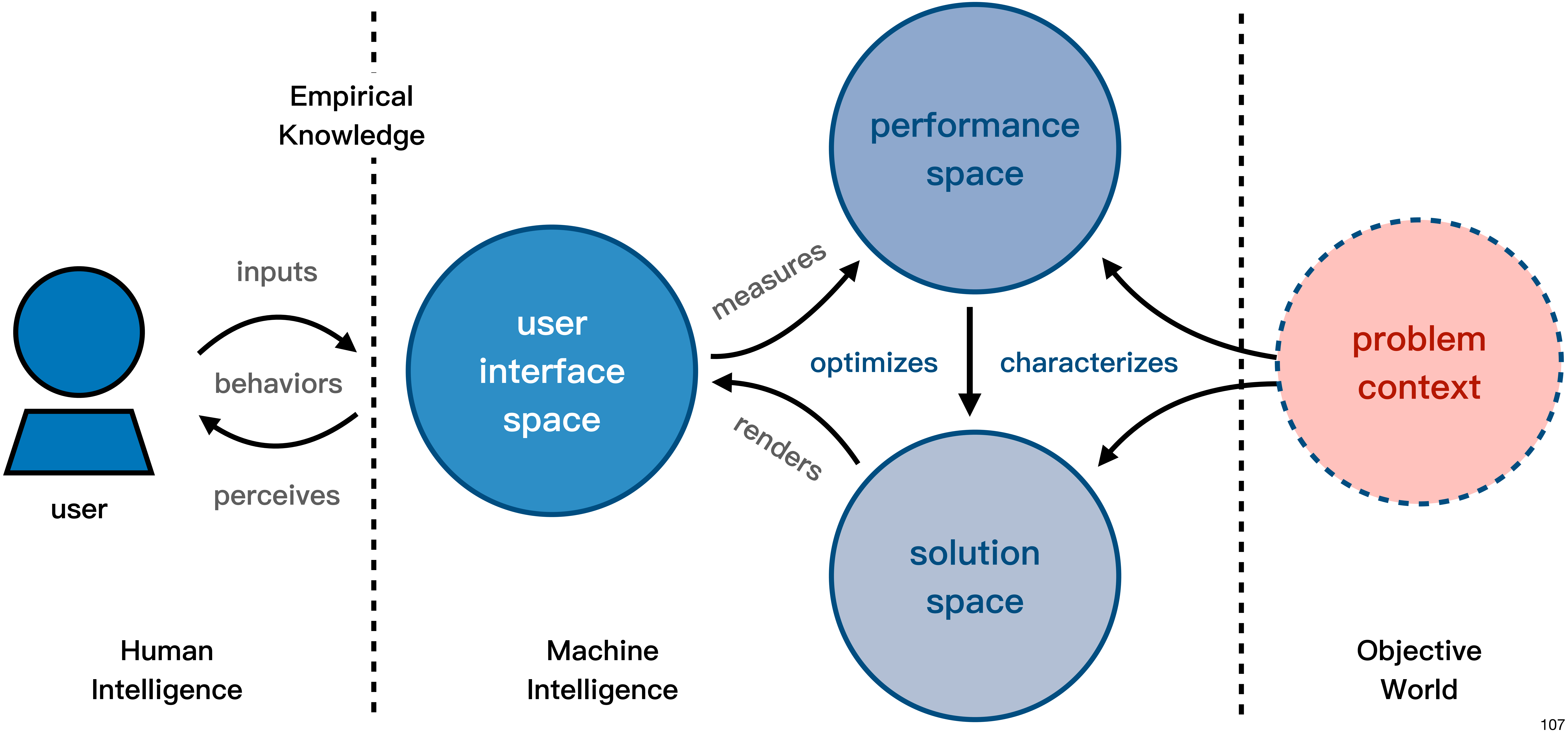
Building Blocks in HITL Optimization Systems



Building Blocks in HITL Optimization Systems



Building Blocks in HITL Optimization Systems



Contributions

Theoretical contribution

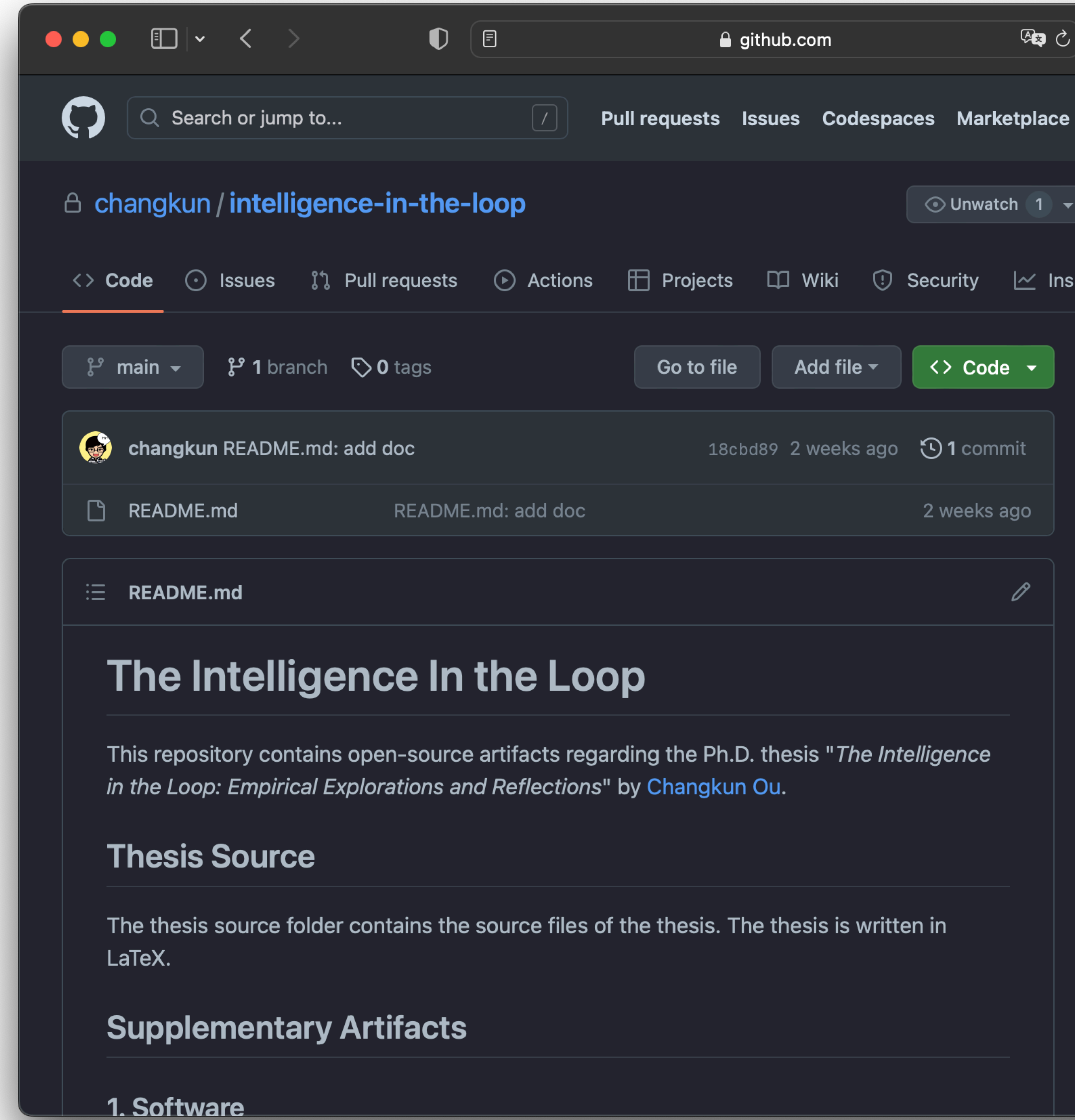
- A framework that analyzes the building blocks of HITL optimization systems
- Interpretation of when HITL is beneficial
- Taxonomy of HITL Input UIs

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

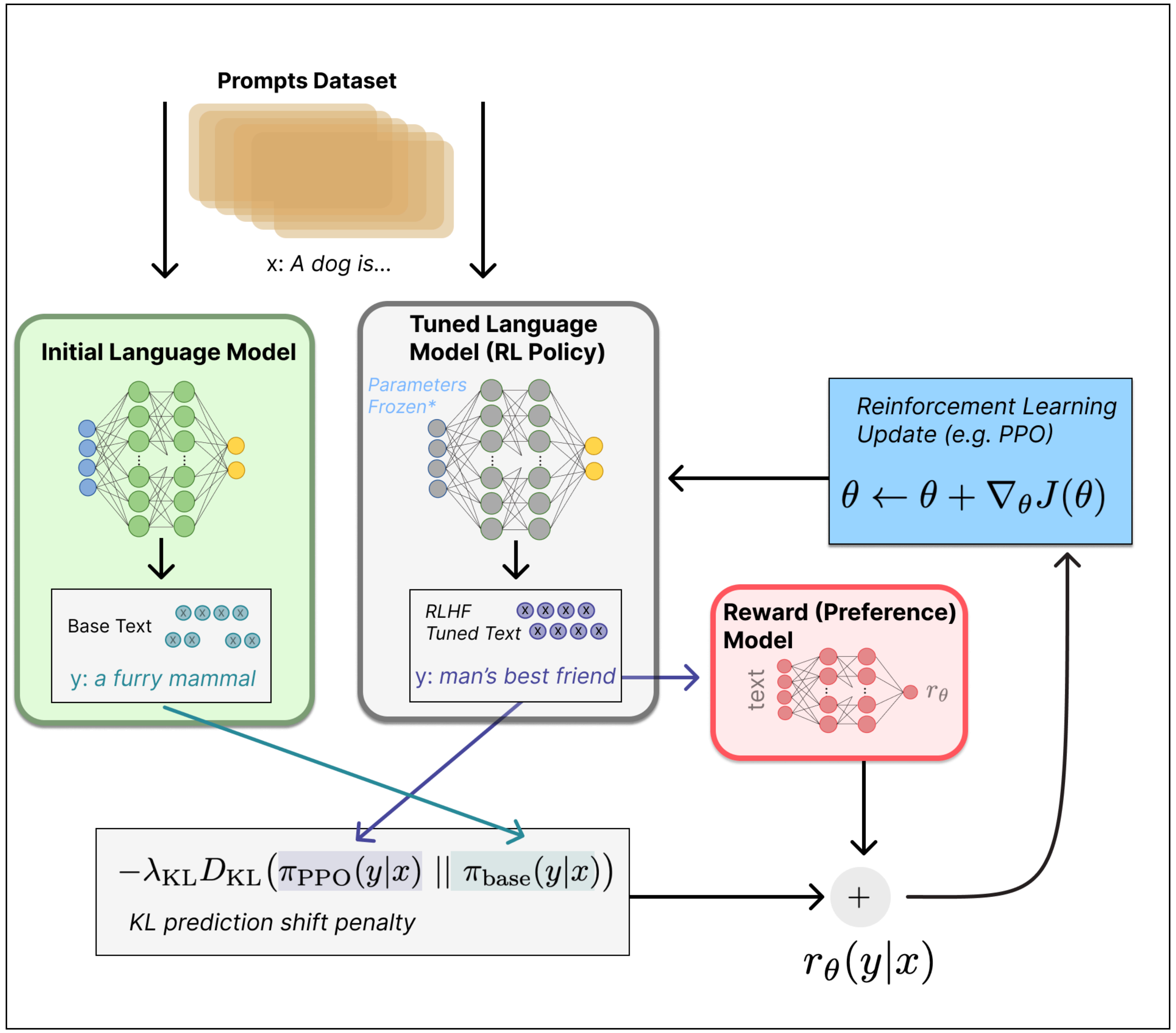


Reflections

- 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

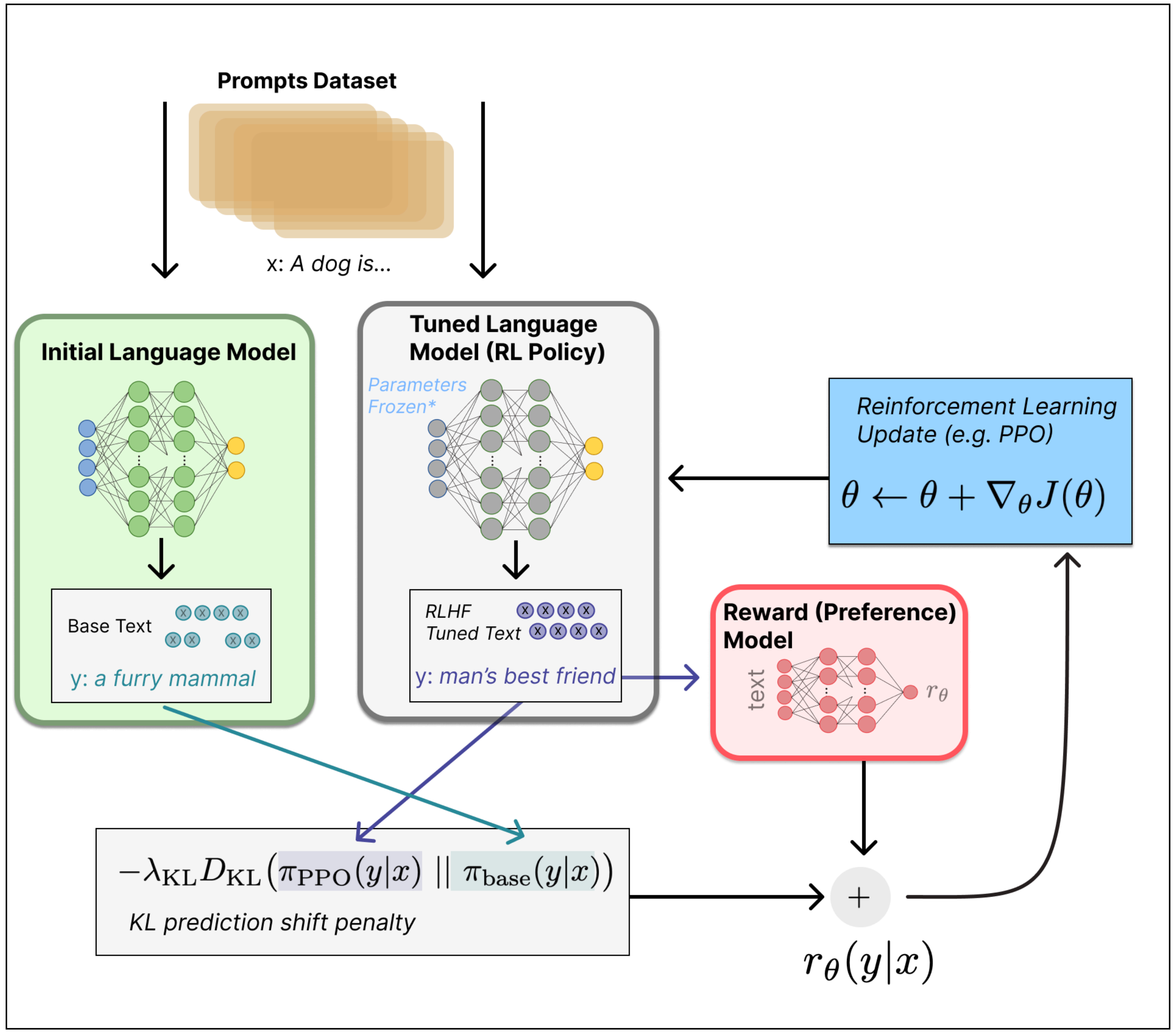
Unstructured and Unaggregated Feedback



[RLHF, HuggingFace, 2023]

- 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

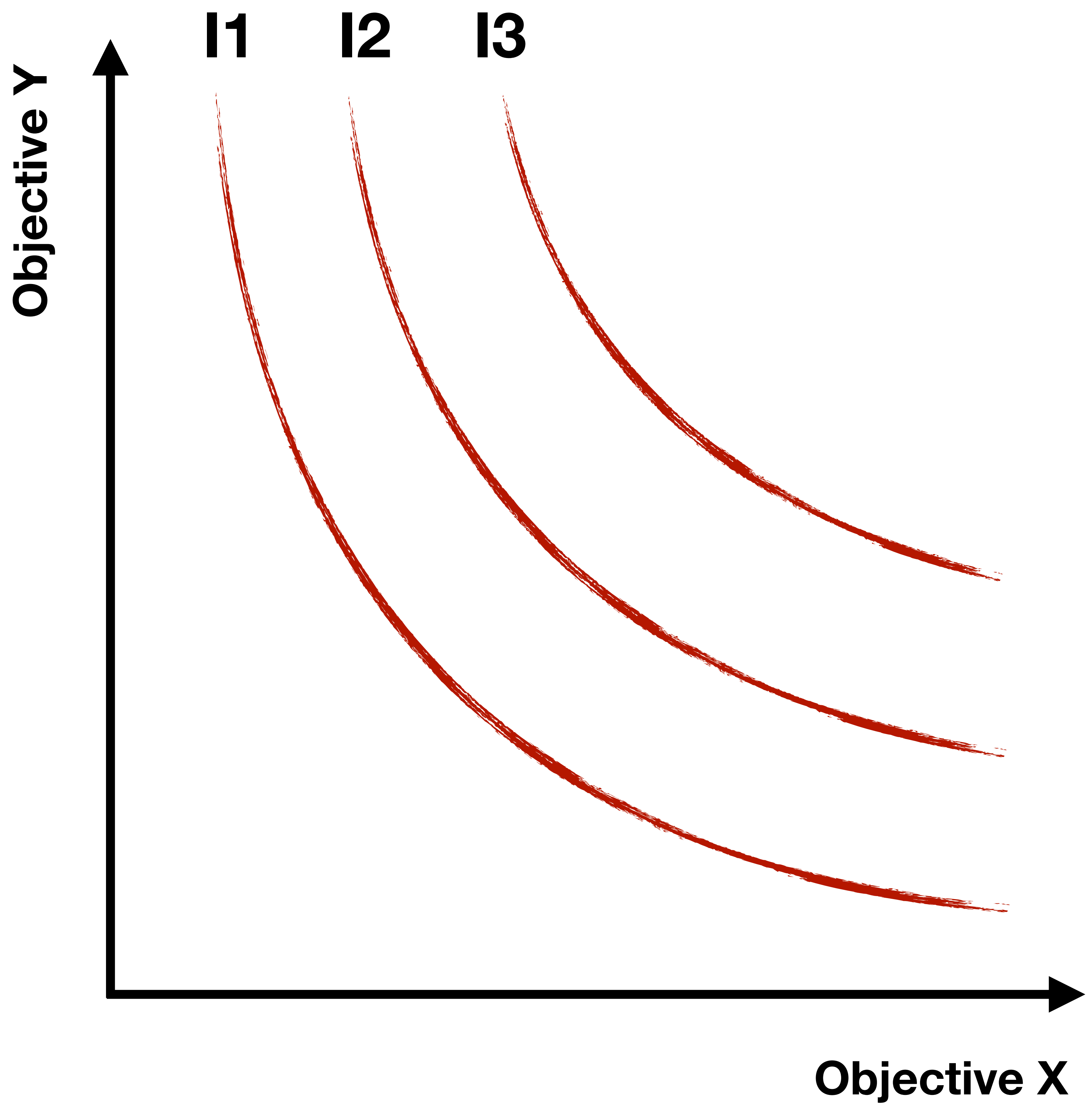


[RLHF, HuggingFace, 2023]

- 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

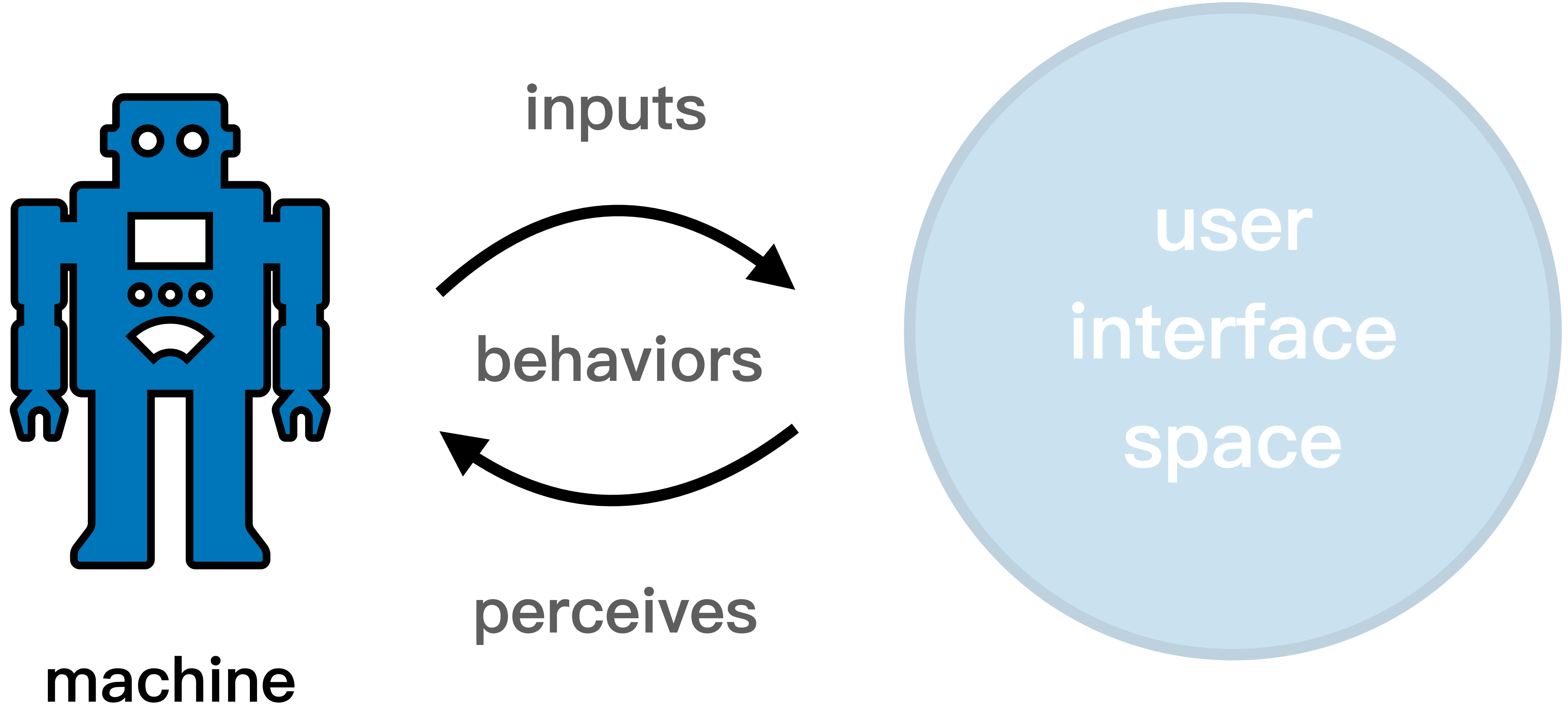


Indifference Curve

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

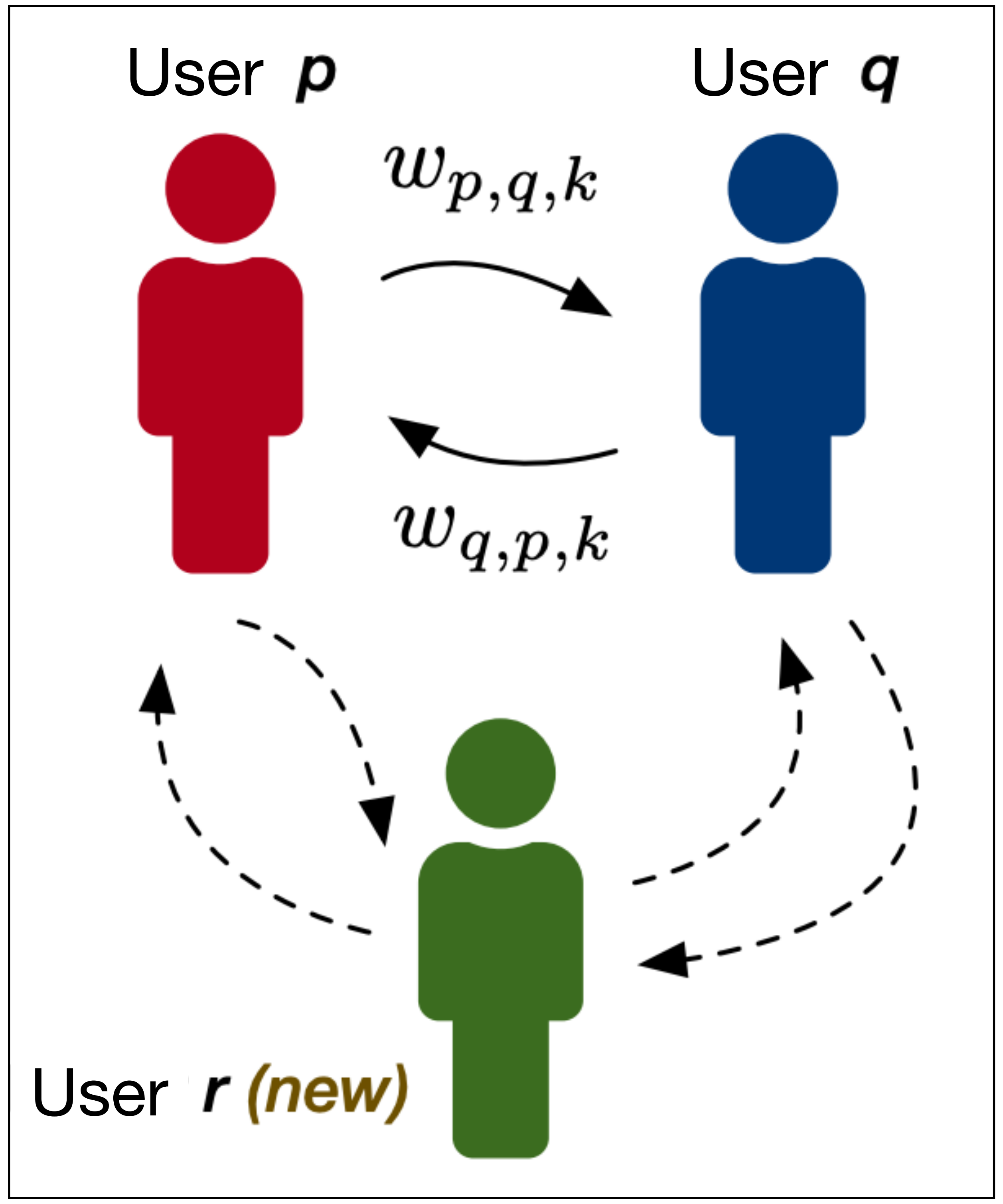
- Modeling indifference and incomplete preference [Nguyen et al. 2021] [Nielsen and Rigotti, 2022]

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]

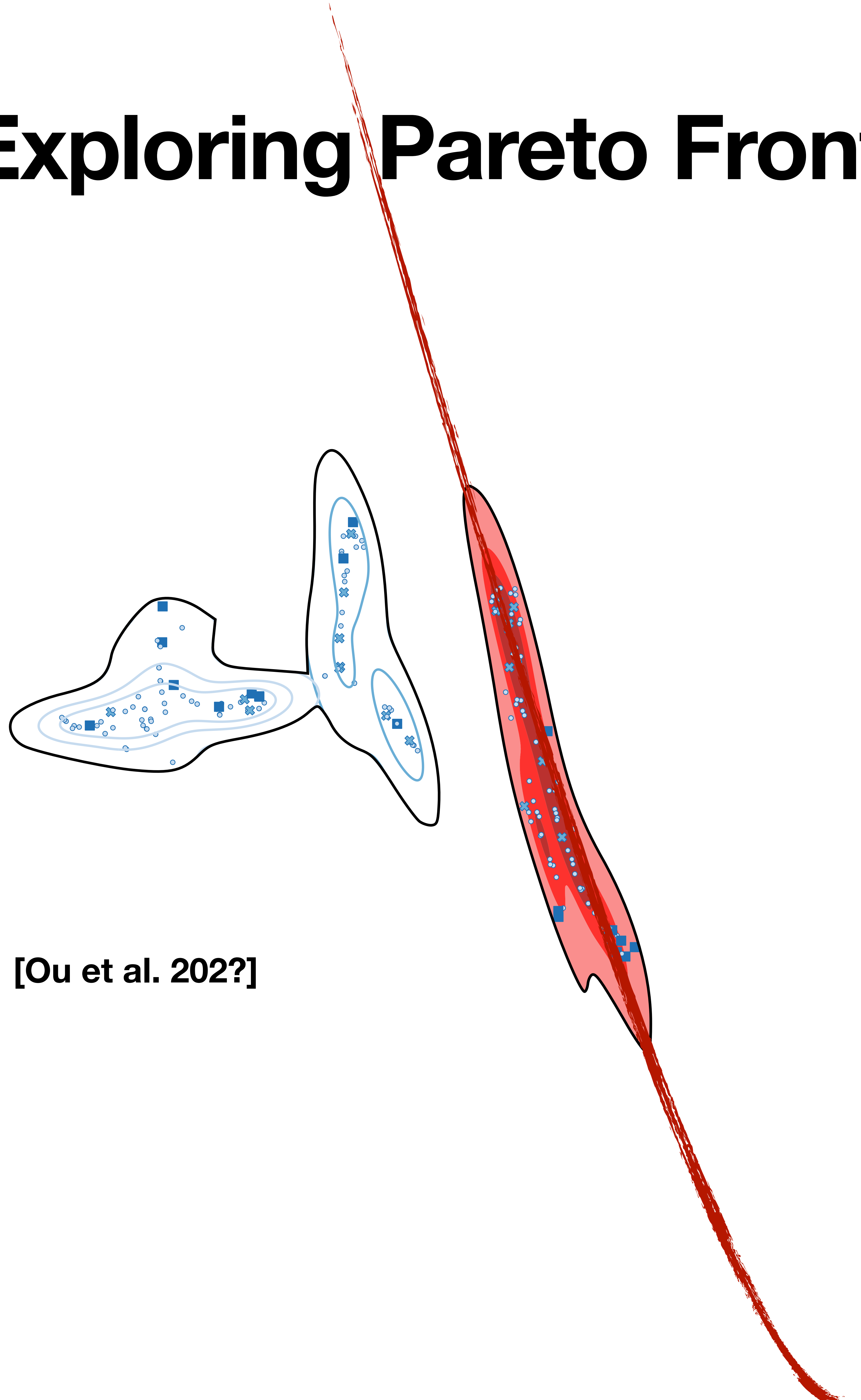
Collective Optimization



[Ou et al. 2019]

- 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]

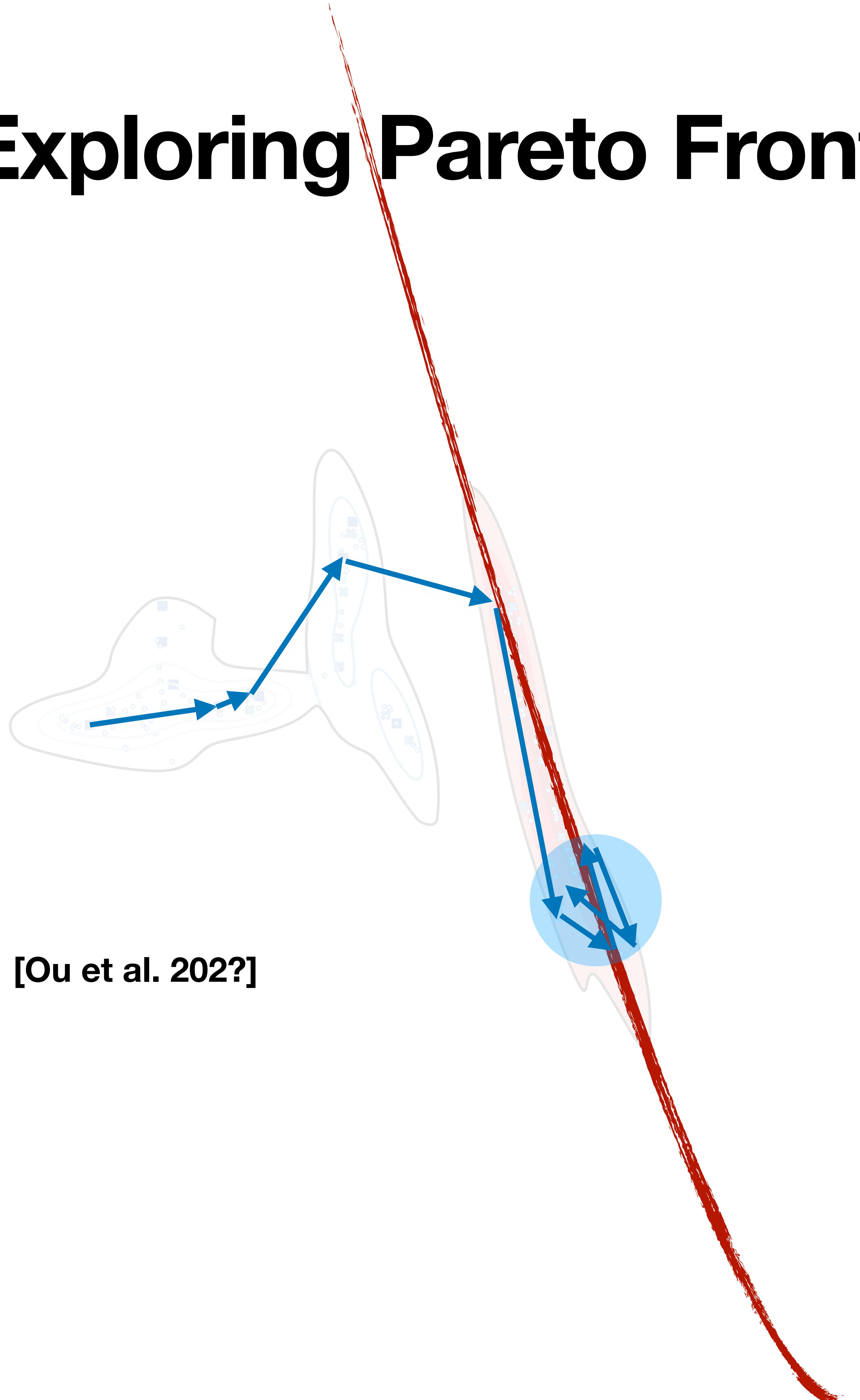
Exploring Pareto Front



[Ou et al. 202?]

- 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

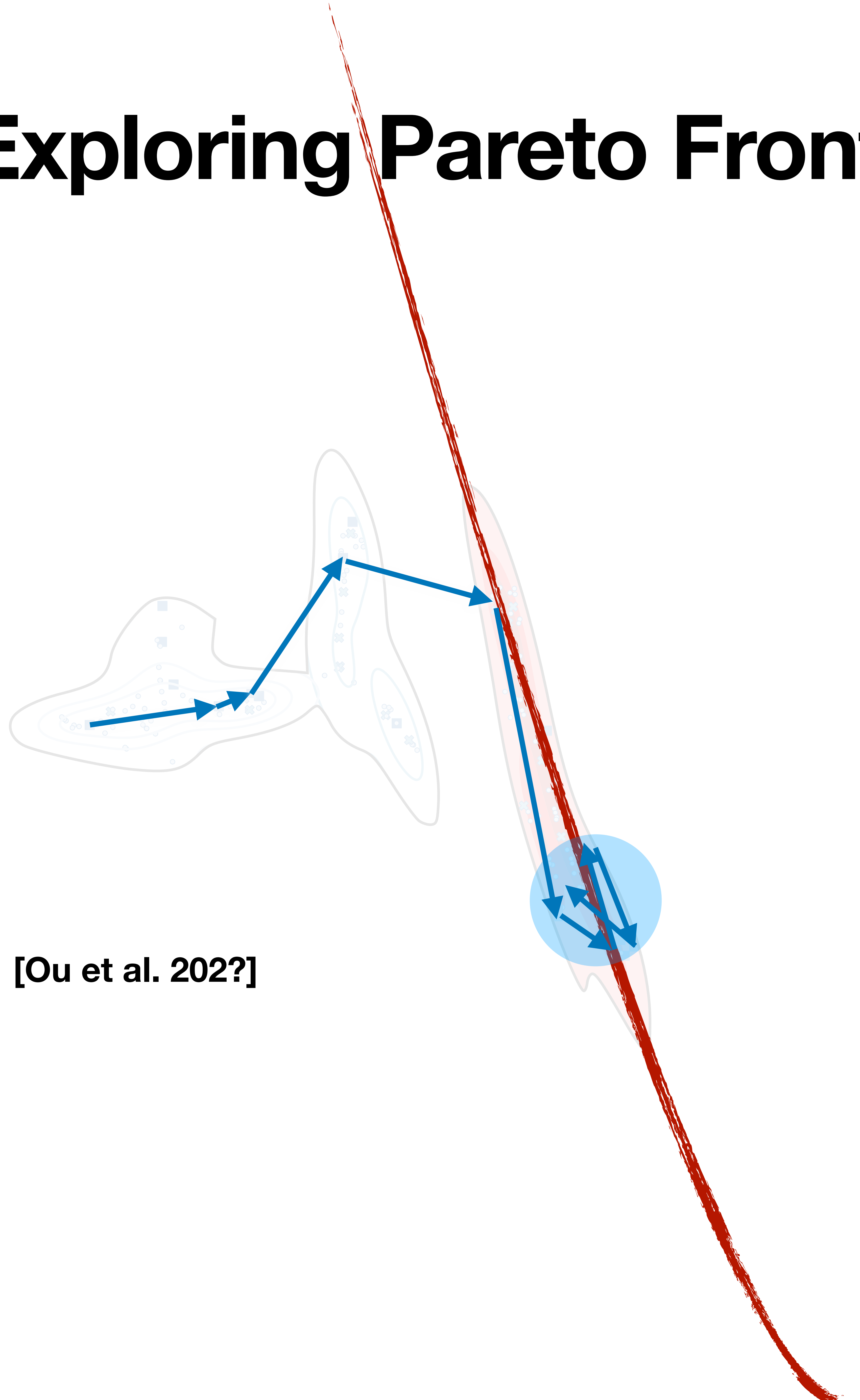
Exploring Pareto Front



[Ou et al. 202?]

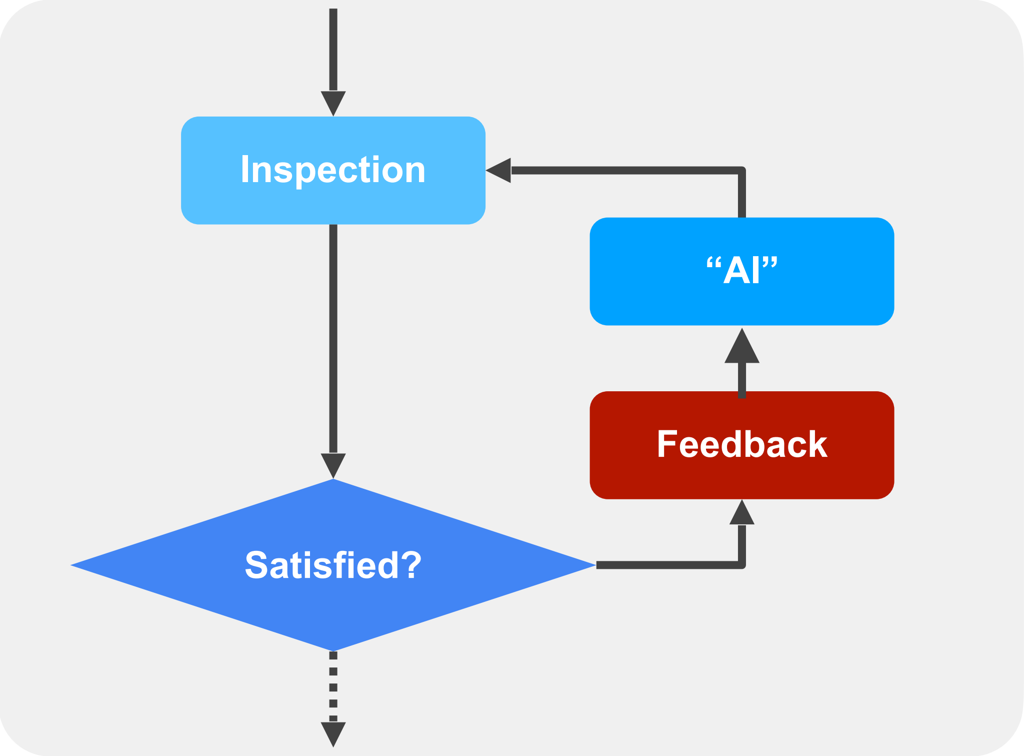
- Involving unstructured and unaggregated human feedback, e.g. reinforcement learning from human feedback using proximal policy optimization [Schulman et al. 2017]
- Modeling indifference and incomplete preference [Nguyen et al. 2021] [Nielsen and Rigotti, 2022]
- Simulating human priors [Murray-Smith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

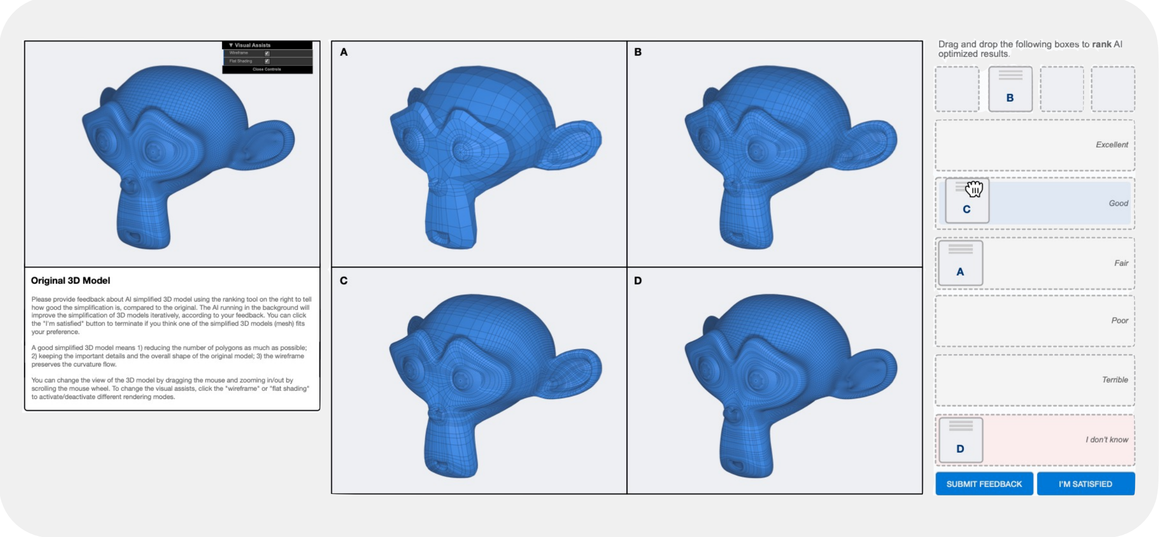
Exploring Pareto Front

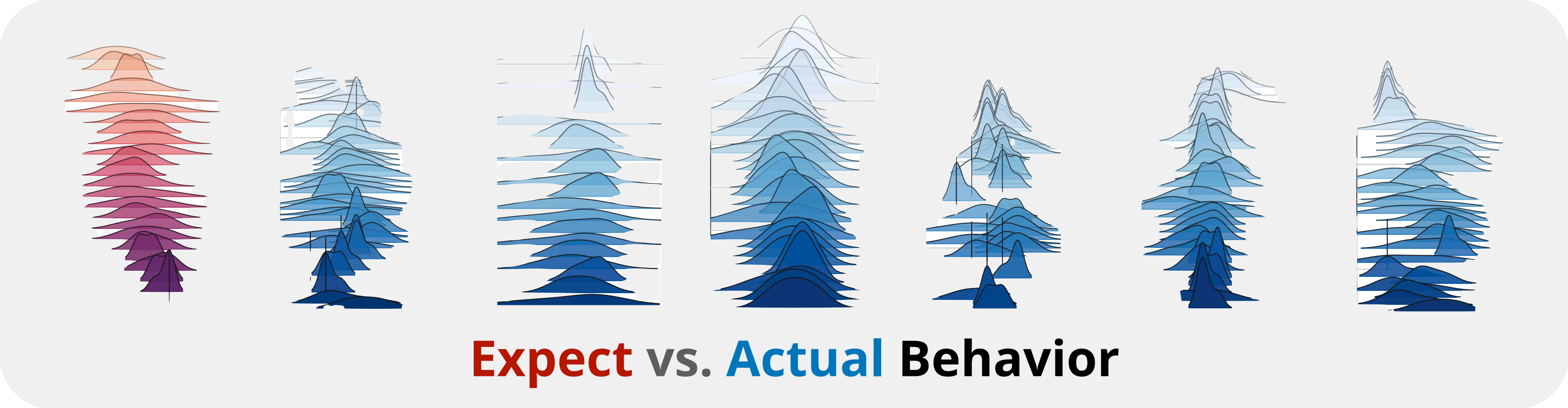


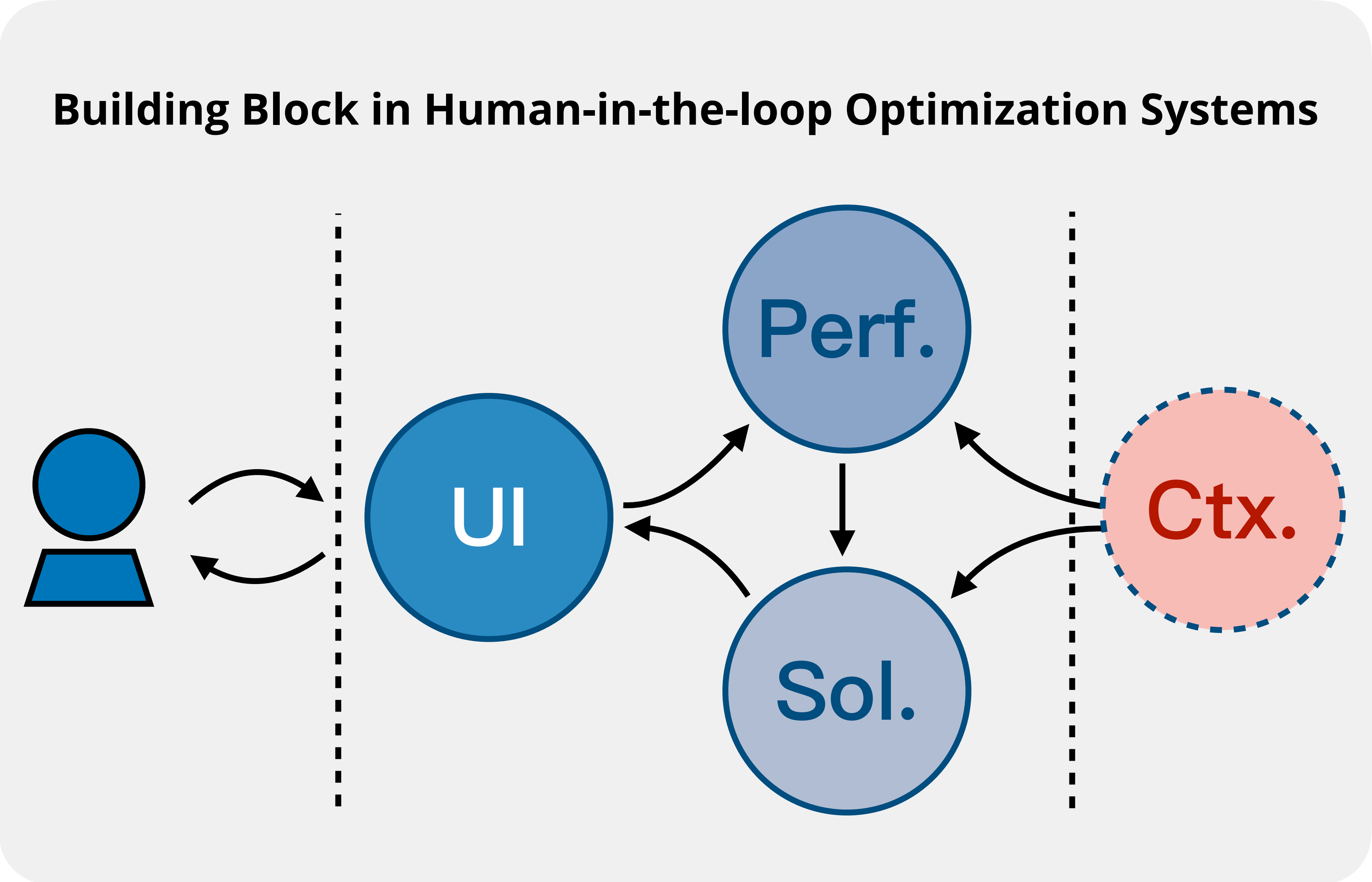
[Ou et al. 202?]

- 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
- ….









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