Relative Design Acquisition: A Computational Approach for Creating Visual Interfaces to Steer User Choices

George B. Mo
gm621@cantab.ac.uk
University of Cambridge
United Kingdom

Per Ola Kristensson
pok21@cam.ac.uk
University of Cambridge
United Kingdom

ABSTRACT

A central objective in computational design is that an optimal design is desired which optimizes a performance metric. We explore a different problem class with a computational approach we call relative design acquisition. As a motivational example, consider a user prompted to make a choice using buttons. One button may have a more visually appealing design and hence is visually optimal to steer users to click it more often than the second button. In such a design case, a relative design is acquired of a certain quality with respect to a reference design to guide a user decision. After mathematically formalizing this problem, we report the results of three experiments that demonstrate the approach's efficacy in generating relative designs in a visual interface preference setting. The relative designs are controllable by a quality factor, which affects both comparative ratings and human decision time between the reference and relative designs.

CCS CONCEPTS

• **Human-centered computing** → **Systems and tools for interaction design**; **Interaction techniques**; **Interaction design process and methods**.

KEYWORDS

Interface Design; Bayesian Optimization; Computational Interaction; Human-in-the-Loop

ACM Reference Format:
George B. Mo and Per Ola Kristensson. 2023. Relative Design Acquisition: A Computational Approach for Creating Visual Interfaces to Steer User Choices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3581028

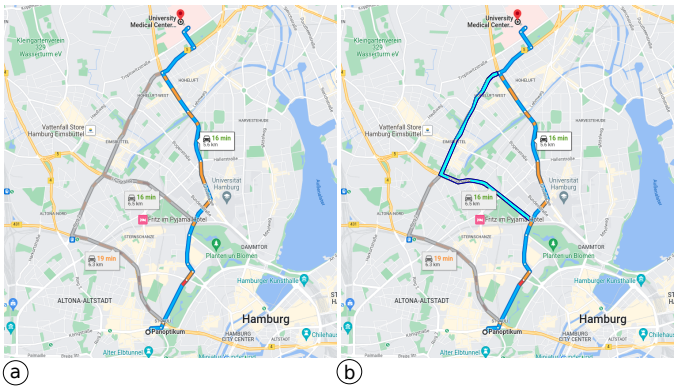

Figure 1: Potential application of relative design acquisition to a maps scenario. Currently in Google Maps, the most desired path is shown in blue whereas the other options are greyed out (Fig. a). However, a secondary preferred option can be designed with relative design acquisition as in (b) to be relative of a certain quality with respect to the reference path design to prompt user path taking.

1 INTRODUCTION

Computational interaction is an emerging field that leverages machine learning methods to aid in the design and implementation of interactive systems and techniques [24]. An overarching objective in computational interaction and in computation design is to optimize a design over its design parameters to maximize a performance metric, such as speed and accuracy [1, 10], ergonomics [34], or user ratings [16, 17]. Approaches taken include using Bayesian optimization for human-in-the-loop interface design [9], neural network based methods [32], combinatorial optimization [19, 23], and linear programming techniques [18]. However, in this paper, we take on a separate class of problems in computational design which instead focuses on finding a *relative design* of a certain *quality* in relation to a reference design.

Consider the following scenario. In a map application (Figure 1a), we are often presented with two choices of route journeys. The visual interfaces for these two choices are designed to entice the user to pick a particular path as to improve the time of travel. However, consider the use case of alleviating the overall traffic in a geographic area. For instance, in this scenario for Google Maps, the more recommended routes could be designed with a bolder border and a brighter color of a controllable visual appeal to the user so as to elicit certain proportions of route traffic selection in Figure 1b. Hence, the routes are designed to be optimal in the sense that they are designed to attract more users to select certain routes whereas other alternative routes are designed to have a relative quality below the desirable route to sway user attention away from selecting it with varying degrees.

From this scenario, it is apparent that with respect to the objective of visual attractiveness and attention-drawing, not only is it desirable to find the optimally attractive interface, it is also desired to identify a relative interface for the second option so as not to

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581028

attract the customer to it, and to do so with a controllable quality. This is a different problem from that of design optimization because instead of searching for a design optimum, we require a *relative design* in relation to a *reference design*. Note that frequently this reference design is an optimal design, or a close approximation of an optimal design.

For many applications, we suggest that a desired characteristic of relative design acquisition is that the relative design's quality can be controlled in relation to a reference design. For example, if the relative button in the web-page was designed to be so poor that the text would not be visible or the button size would be too small, this would render the button completely inaccessible and hence would not satisfy basic established web design principles. This would be an undesirable result. To avoid this it is necessary that the relative design is *controllable* by some factor and we call this design parameter *quality*. This design parameter is essential for allowing the designer to tune a relative design.

In addition, another desirable characteristic of relative design acquisition is that this design technique is applicable and demonstrably effective in synthesizing relative designs in relation to a reference design based on user feedback. In the maps example, a method in which attractiveness can be assessed is through the rating or reaction of the user to the design. Therefore, another desirable characteristic of relative design acquisition is that the proposed method is effective in a human-in-the-loop setting, as in based upon iterative user feedback, and generates new and more fitting designs. Further, the generated relative designs should be consistent in quality for an entire population of users, not just individualized designs for specific users. This allows the relative design acquisition technique to be more widely applicable for creating relative designs that are effective for groups of individuals with respect to some desired outcome or objective.

To attempt to tackle the problem of relative design acquisition, as motivated from the example above, the rest of this paper presents a computational method that incorporates the above desired characteristics and reports the results of three experiments that demonstrate the efficacy of this approach.

In summary, the main contributions of this work are:

- A mathematical formalization of the problem of relative design acquisition and a generalized framework for computing quality-controllable relative designs using a computational approach.
- An empirical investigation of relative design acquisition for visual context-free and visual context-present scenarios for individual user preferences in human-in-the-loop experiments.
- An evaluation to understand the performance and potential uses of relative design acquisition applied to creating reference and relative visual designs for a new population of users in a shopping website design context example, highlighting its effectiveness in generating relative designs and effects on human decision time.

In this paper, we primarily focus on the application of relative design acquisition for button interfaces in the context of a shopping website (Figure 2). On such a website, users are often presented with two choices—to either purchase the product at hand, or not to

Figure 2: Example of two buttons adjacent on the bottom right that lead to different options of buying using Amazon Prime (*Join Prime Today* versus *Continue with Free One-Day Delivery*). The button urging customers to buy is designed to be more visually attractive than the one that allows the customer to continue without buying. The two buttons are termed as *reference* and *relative* respectively.

purchase it. The visual interfaces for these two choices are designed to entice the customer to pick the button to purchase more. For instance, in this scenario for Amazon Prime, the button to buy the product is designed with a brighter yellow color with a more discernible border and darker font color compared to the "continue without buying" option. Relative design acquisition can conceivably be applied to other scenarios, but this application was chosen to be explored in depth to account for the many differing factors that could affect user behavior. Specifically, the first study in this paper explores the effects of memorability and the lack of visual context. The second study investigates the effects of a visual context and the sampling method. Finally, the third study examines the effects of different tasks on ratings and decision times for relative designs generated for a new set of participants.

We also draw attention to the fact that this specific shopping website example could lead to ethical concerns on its use as a dark design pattern for user manipulation, which we specifically address in Section 4.1. We stress that dark designs could be an easy application of this method, and emphasize the cautionary approach that would have to be taken when using relative design acquisition, that is, aligning the intent of steering user choices with ethical considerations.

Through exploring the different effects of applying relative design acquisition to visual interfaces, our method shows promise in generalizing to many scenarios where multiple choices are presented and users should be nudged to make a certain choice over the other. For instance in guiding a user around a tutorial for a productivity application, less frequently used or less useful tools can be designed to be suboptimal in their visual affordance. A further example would be training wheels for a user interface [5], where

the interface designs of advanced features can be gradually adjusted in quality to become more prominent as users gain proficiency. The key advantage of our method is that it can generate systematically different design candidates of controllable quality to steer user choices. Note that in all the above applications, controllability in guiding user decisions is especially important to yield controllable user choice proportions for a certain objective, such as to alleviate overall traffic in an area for a map application.

2 APPROACH

2.1 Background

To introduce the method for relative design acquisition, we first detail Gaussian Processes and how they can be applied in human-in-the-loop applications. For clarity, we note that our method, relative design acquisition, is based on Gaussian Processes only and the assumption that we can build a Gaussian Process model based on data we collect from the user. We emphasize that relative design acquisition tries to find a relative design of a certain quality with respect to a reference design. It is independent, and not reliant on, the way the data is gathered or sampled to build the Gaussian Process model, and is a generalized method for generating a relative design given a reference design.

2.1.1 Gaussian Processes. For many applications, the relationship between the input parameters $\mathbf{x}$ and the output from function $f : \mathcal{X} \rightarrow \mathbb{R}$ is unknown, which in our context is a design objective function. This for instance could be $\mathbf{x}$ representing the configuration for a particular design and f be the user performance to that design, such as speed, accuracy, or user rating. We can model f as a sample from a Gaussian Process (GP), which is a collection of dependent random variables for each $\mathbf{x} \in \mathcal{X}$, where every subset is distributed as a multivariable Gaussian. A Gaussian Process $GP(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ is specified by its mean function $\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and kernel function $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))]$. Let the GP prior be $GP(0, k(\mathbf{x}, \mathbf{x}'))$.

Suppose we sample $\mathcal{X}$ at points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ and retrieve $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, where $y_i = f(\mathbf{x}_i) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. Gaussian noise. Then, the posterior over f is a GP distribution with mean $\mu(\mathbf{x})$, covariance $K(\mathbf{x}, \mathbf{x}')$, and variance $\sigma^2(\mathbf{x})$:

$$\mu(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}$$
$$K(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}')$$
$$\sigma^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x})$$

Here, $\mathbf{k}(\mathbf{x}) = [k(\mathbf{x}_1, \mathbf{x}), ..., k(\mathbf{x}_n, \mathbf{x})]^T$ and $\mathbf{K} = [k(\mathbf{x}, \mathbf{x}')]_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}}$. Notably, we use GPs to create a surrogate model of the relationship between the design parameters $\mathbf{x} \in \mathcal{X}$ and the design objective function f representing the user's performance on that design parameter. As f is a black-box system, this is an appropriate model as user performance over design parameters is assumed to be smooth and continuous across $\mathcal{X}$. In human-computer interaction, Gaussian Processes have been applied in many computational design scenarios to optimize designs (e.g. [9, 16, 17]), due to its ability to accurately model user behavior given limited data and its low computational overhead for human-in-the-loop applications.

2.2 Computational Approach to Relative Design Acquisition

2.2.1 Approach for Controllability. We now detail the novel mathematical formulation of the problem we here term *relative design acquisition*. This method is distinct from, and modular to, the sampling method from which the data is collected for creating the Gaussian Process. Suppose that the design problem we are working on involves a black-box system design objective function that takes design parameters as inputs and outputs an objective value, namely $f : \mathcal{X} \rightarrow \mathbb{R}$. Suppose we have decided upon a reference design $\mathbf{x}^*$ from which we want to find a relative design. Much of the time, $\mathbf{x}^*$ would be an optimal design with respect to f, and has an objective function value of $f(\mathbf{x}^*)$. Suppose that we have also collected data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ where $\mathbf{x}_i$ are the design parameters and y_i are the corresponding objective function observations (i.e. $y_i = f(\mathbf{x}_i) + \epsilon$, where ϵ is some noise term). We further assume the data $\mathcal{D}$ is then used to create a surrogate model for $f(\mathbf{x})$ using a GP.

From the GP we can infer the predictive mean of $f(\mathbf{x}^*)$. Say it is μ. Then say for the relative design, we want to find one that is γ in quality compared to the reference design $\mathbf{x}^*$. That means that we want to find a relative design $\mathbf{x}^a$ such that $\mathbb{E}[f(\mathbf{x}^a)] = \gamma(\mu - f_l) + f_l$, where f_l is the lower bound of the function f.

This formulation allows us to construct an objective function to find the relative design $\mathbf{x}^a$ with quality γ. Essentially, we want to find a relative design $\mathbf{x}^a$ such that $\mathbb{E}[|f(\mathbf{x}^a) - (\gamma(\mu - f_l) + f_l)|]$ is minimized. Hence, our objective function for the relative design is:

$$RDA(\mathbf{x}^a | \mathbf{x}^*) = \mathbb{E}_{f(\mathbf{x}^a)}[|f(\mathbf{x}^a) - (\gamma(\mu - f_l) + f_l)|] \quad (1)$$

We now explain more details about this objective function, especially the role of γ. We will call the objective function in equation 1 the *RDA objective*, which differentiates it from the design objective function to which we evaluate a design. As in many of our applications, μ represents the objective function value or performance of the reference design. As the reference design is frequently some sort of optimal design, we limit γ to be $0 \leq \gamma \leq 1$. For example, if $\gamma = 0.9$ in the RDA objective, then $\mathbf{x}^a$ should have a design objective function value that is 90% of that of the optima, and hence the two designs—relative and reference—should have approximately the same quality. However, if $\gamma = 0.5$, then $\mathbf{x}^a$ would have a design objective function value that is half of that of the reference design, and hence the relative and reference designs should have a greater discrepancy in observed quality.

Note that we can actually replace μ with any observation of the objective value y we get that corresponds to the reference design we want to find a relative one to. It does not have to be the mean of the predictive distribution of $f(\mathbf{x}^*)$, but can simply be an objective value that corresponds to the performance of $\mathbf{x}^*$.

Appendix A shows an explicit expansion of the RDA objective into an analytic form given a GP model. We stress that relative design acquisition is not an iterative optimization process and it does not aim to find an optimal design to maximize the design objective function. We expect that, for some applications, optimizing the RDA objective will yield several relative design candidates of a certain quality with respect to a reference design. In such a case, the designer will have to use design judgment or user studies to choose a preferred generated relative design.

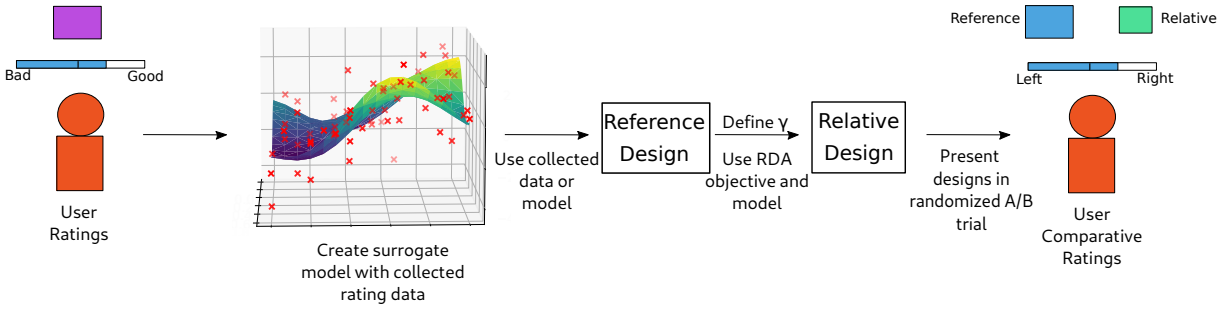

Figure 3: Computational framework for relative design acquisition, where the data collected from the users from an arbitrary sampling method is then used to generate the reference design and the relative design. For the relative design, it is obtained through optimizing the RDA objective and using the surrogate model. The two choices are then presented to the user for a comparative rating.

2.2.2 Integrating with Different Sampling Methods. As may be realized from the previous subsection, the sampling method to build up the data $\mathcal{D}$ to create the GP is something that does not need to be fixed, and hence relative design acquisition can be integrated with many different types of data collection settings. Thus, the sampling method (which could be random sampling, Bayesian optimization, etc.) and relative design acquisition are modular. We describe below how random sampling and Bayesian optimization can be integrated as sampling methods to obtain a reference design from which relative design acquisition can be used to generate the relative design.

- *Random sampling* can be used to generate a dataset $\mathcal{D}$. The surrogate model (GP) can then be built using $\mathcal{D}$. To find the reference design that is near-optimal, one option is optimizing over the posterior mean of the GP in $\mathcal{X}$, namely retrieving $\mathbf{x}^*$ and $\mu = \mathbb{E}[f(\mathbf{x}^*)]$. Using $\mathbf{x}^*$ and μ or the achieved objective value $y^* = f(\mathbf{x}^*) + \epsilon$, and selecting γ, we can use relative design acquisition to find the relative design $\mathbf{x}^a$.
- *Bayesian optimization* allows an optimum design to be searched for the black-box system design objective f and can be thought of as an iterative sampling method. As a result of iterative sampling, a dataset $\mathcal{D}$ is created from which a surrogate model of f can be generated using a GP. The surrogate model and relative design acquisition can then be used to find a relative design $\mathbf{x}^a$ of which a reference design $\mathbf{x}^*$ (the maximum we have seen thus far) is compared to. The design objective value of the reference point x^* can be taken as either the posterior mean in $\mu = \mathbb{E}[f(\mathbf{x}^*)]$ or the achieved design objective value y^* corresponding to $\mathbf{x}^*$.

In summary, relative design acquisition can be integrated with a wide variety of sampling techniques and settings. The complete framework for relative design acquisition is shown in Figure 3.

2.2.3 Practical Considerations. In all our experiments, we normalize each of the design parameters so that they each lie within the range $[0, 1]$, and each of the optimizations of the RDA objective is done with the optimization algorithm L-BFGS-B with 25 restarts. Also, the design objective function is normalized so that it is within the range $[-1, 1]$, hence $f_l = -1$. The GP used has a radial-basis function kernel with length-scale l, composed with a white noise kernel with a noise level σ. In addition, when new data is added, the hyperparameters $\{l, \sigma\}$ of the GP are also updated by maximizing the maximum log likelihood.

3 EXPERIMENTS

To understand how relative design acquisition can be applied to practical interactive design scenarios, we conduct three different studies all of which involve the design of visual interfaces. The objective function used to assess the quality of the generated visual interfaces is the absolute user rating of the designs. Study 1 focuses on a visual context-free scenario, Study 2 focuses on a visual context-present scenario of a shopping website, and Study 3 focuses on evaluating the generated relative designs with a population of users. All three studies were performed on a ASUS ProArt Display monitor, 1920 × 1200, using a keyboard and mouse. The studies were approved by the local ethics committee and in all studies, participants were not told about the concepts of reference and relative designs. The only inclusion criteria in all studies was that the participant is not color-blind.

3.1 Study 1: Visual Context-Free Interface Design

3.1.1 Goals. The primary goal of the first study is to understand if the method of relative design acquisition can be applied to find relative designs in a visual context-free scenario for individual users. We mean visual context-free as in only the interface is displayed without any other visual context. Specifically, we want to explore the following research questions:

(1) Can relative design acquisition capture individual visual preferences of users in a visual context-free scenario, i.e. do people prefer the relative designs less?
(2) If so, do the generated relative designs capture individual visual preferences that do not change when sampled various times for the same user?
(3) Can relative design acquisition be controlled by the design parameter γ (the *quality*)?

(4) Does the amount of data collected impact the quality of the relative designs generated?

3.1.2 Participants. To address the above questions, we conducted an experiment as a within-subjects design. We recruited 16 participants (12 males, 4 females) with an average age of 21.7 ($sd = 1.3$). The participants were recruited through opportunity sampling within our institution. None of the participants had any visual impairments.

3.1.3 Task. In the study participants were asked to rate a button or text style generated from Bayesian Optimization or relative design acquisition on the basis of whether they think the design would attract customers in a hypothetical scenario of a shopping website.

We used two different rating tasks: 1) an *absolute rating task* in which the participant was asked to indicate whether a generated button or text style was bad or good using a continuous rating scale; and 2) a *comparative rating task* in which a participant was presented with a reference design and a relative design positioned next to each other in a randomized A/B test and asked to manipulate a continuous slider towards the design that he or she thought was better. An example of an absolute rating task in Study 1 is shown in Figure 4a and an example of a comparative rating task is shown in Figure 4c.

Both rating tasks (absolute and comparative) used a continuous scale, however, the precise continuous value was not shown to participants. The internal scale was from zero to ten. For the comparative rating task we chose a continuous comparative rating to evaluate the effect of γ on the preference in a continuous manner as opposed to a discrete manner (i.e. selecting one button over the other).

In the text task, to change the font style of the text displayed, we take the Adaptifont text generation process as detailed in Kadner et al. [13]. They took candidate fonts and performed PCA to project the font styles onto three distinct axes from which continuous style changes in fonts can be generated. The design parameters and corresponding parameter ranges in the text task were:

- Font size – [15.0, 40,0]
- Transparency – [0.1, 1.0]
- Font axis 1 from Adaptifont – [3.0, 13.0]
- Font axis 2 from Adaptifont – [3.0, 13.0]
- Font axis 3 from Adaptifont – [3.0, 13.0]

The three axes corresponding to the font style of the text from Adaptifont [13], roughly control the horizontal and vertical scaling, and the serif. The range for the font size was selected so that the text would be at an appropriate size for a text link on a web page. The range for the transparency was selected so that the text would be visible, and the three font axes ranges were selected to allow the text to be successfully rendered.

In the button rating task, the absolute rating task and comparative rating task layouts were very similar to the text task. The design parameters in the button task were:

- Size – [0.5, 2.0]
- Transparency – [0.1, 1.0]
- Hue of the button colour – [0.0, 1.0]
- Saturation of the button colour – [0.0, 1.0]
- Value of the button colour – [0.0, 1.0]
- Width of the button border – [0.0, 1.0]

The rating interfaces for the button task are shown in Figures 4b and d. Specifically, the ranges for the size and transparency were selected for both tasks so that they could be realistically rendered on a web-page and be seen by the user. The upper bound for the border width was selected to make it appropriately thick for the smallest button size. To clarify, for both tasks the black-box design objective function would be the user rating of the design, and we chose to make the text the same for the text rating task so that the two designs are semantically the same.

3.1.4 Procedure. For the study protocol, we have 2×2 conditions: 1) the specific interface element: either a button or text; and 2) the quality design parameter: generating the relative designs based on either $\gamma = 0.6$ or $\gamma = 0.9$. The order of all conditions were counterbalanced across the 16 participants. Participants were given the prompt of judging the button or text based on the quality of whether they thought it would attract the most customers on a shopping website using their imagination. Each condition involved two phases separated by a 3 minute break - 1) the Bayesian Optimization phase that involved both absolute and comparative rating tasks; and 2) the comparative rating phase only involving comparative rating tasks.

We started each condition with a tutorial describing the task followed by 10 absolute ratings and 1 comparative rating to calibrate their preferences and exposure to different designs. The Bayesian Optimization phase was then initialized with 5 random samples for absolute ratings. Then, after every 15 iterations of Bayesian Optimization (absolute ratings), we took the top-3 designs with the highest absolute ratings in the previous 20 iterations. We computed relative designs based on those best designs given the data observed so far (yielding 5 groups per condition). Next, we presented the relative designs with the corresponding reference designs in comparative rating tasks. The above was then repeated for 5 batches, so in total each setting underwent 75 iterations of Bayesian Optimization (absolute ratings) and $3 \times 5 = 15$ A/B comparisons involving relative designs (comparative ratings). For further clarity, Figure 5 shows the procedure of this study, with the tutorial, the Bayesian Optimization phase consisting of 5 groups, and the comparative rating phase.

The participants were then asked to take a 3 minute break. In the comparative rating phase, the $3 \times 5 = 15$ comparative rating tasks generated through the Bayesian Optimization process were then presented in a randomized order to the participant again. This was to see if the relative designs generated for each participant were actually representative of their true personal preferences, and not just an artifact of memorability.

The study finished with a short post-study questionnaire. In the questionnaire, participants were first asked on what basis of quality did they judge the two different design scenarios. They were then presented with some of the button and text designs they had seen during the study, and then asked if what they had answered previously is consistent with what they observed they preferred in the designs. Each participant was compensated £10 for their participation and the entire session lasted approximately one hour.

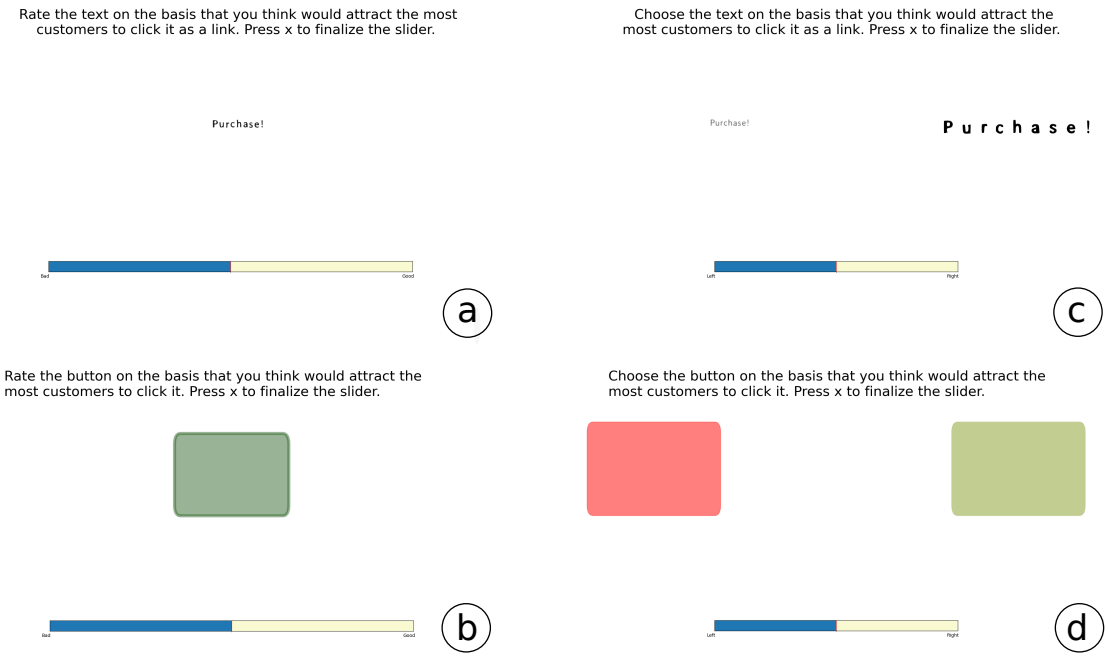

Figure 4: (a) and (b) show the absolute rating interface for the text and button tasks respectively. (c) and (d) show the comparative rating interface for the text and button tasks respectively.

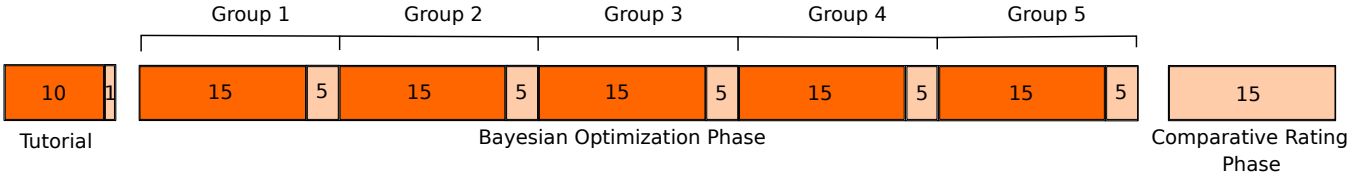

Figure 5: The procedure of Study 1 for each of the four tested conditions, showing the tutorial, the Bayesian Optimization phase, and the comparative rating phase. The darker orange color represents an absolute rating and the lighter orange color represents a comparative rating. The numbers within the boxes show the number of trials performed.

Table 1: Mean and standard variations for the text and button tasks for the two γ and time conditions.

	$\gamma = 0.6$				$\gamma = 0.9$			
	During		After		During		After	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Text	2.426	1.370	2.576	1.160	4.377	1.822	4.315	1.600
Button	3.197	1.981	3.359	1.601	4.577	2.205	4.584	1.795

3.1.5 Quantitative Results. We first analyze the distribution of comparative ratings for the designs generated by relative design acquisition. We term the comparative ratings done during the optimization as "during" and comparative ratings after the three minutes of a break as "after", and term these collectively as the time factor. For each comparative rating, we adjust the rating such that a comparative rating of 10 always indicates a maximum preference of a generated relative design, a rating of 0 always indicates a maximum preference of a generated reference design, and a rating of 5 represents the case when participants cannot really tell the difference between the generate reference or relative designs.

Figures 6a and b show the box-and-whiskers plot for both the text and button tasks' comparative ratings. Table 1 shows the means and standard deviations for each of the 2×2 conditions for both the text and button tasks. A repeated measures two-way ANOVA was

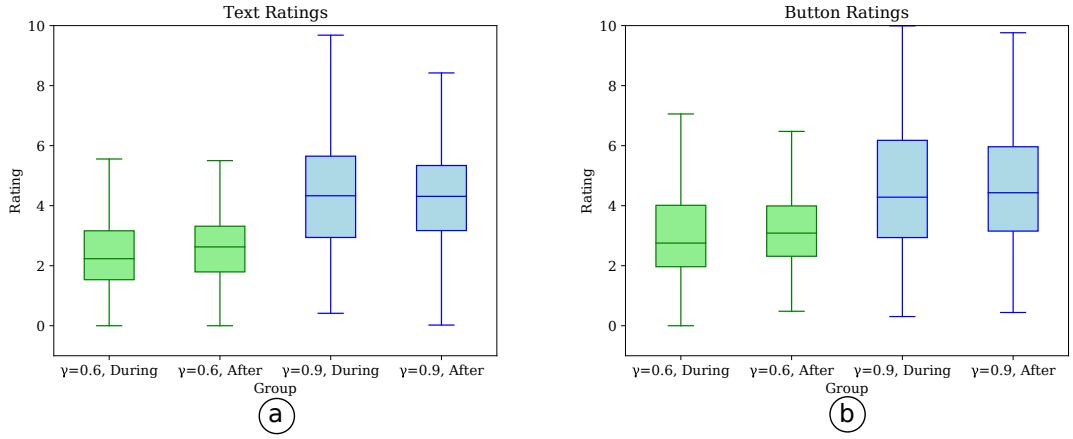

Figure 6: Comparative rating box-and-whiskers plot for (a) the text task and (b) the button task for each condition, with the whiskers representing the first or third quartile ±1.5 IQR.

performed with the within-subjects fixed factors of γ and time (during and after) for both the text and buttons task with the dependent variable of the comparative rating. None of the analyses violated sphericity by Mauchly's test. For the text task, γ was statistically significant ($F_{1,15} = 90.42, p < 0.0001, \eta_P^2 = 0.858$), but the time factor (during vs. after) was not statistically significant in determining the user rating ($F_{1,15} = 0.196, p = 0.664, \eta_P^2 = 0.0129$). For the button task, γ was statistically significant ($F_{1,15} = 19.46, p = 0.000506, \eta_P^2 = 0.565$), but the time factor was not statistically significant in determining the user rating ($F_{1,15} = 1.166, p = 0.297, \eta_P^2 = 0.0721$).

We analyzed the comparative ratings for the 5 groups of the number of data points used to generate the relative designs for $\gamma = 0.6$ (as there were $3 \times 5 = 15$ comparative ratings during Bayesian Optimization) for both the "during" and "after" scores. Note that the data points collected were used to create the GP model from which relative design acquisition was performed. A repeated measures 2-way ANOVA with the fixed within subjects factors of time (during and after) as well as the 5 different data conditions was performed with the dependent variable of comparative rating. For all conditions, there was no significant difference in the data groups, indicating that data may not significantly affect the relative design quality. We also did not observe any statistical significance for the same analyses applied to $\gamma = 0.9$.

3.1.6 Questionnaire Results. All participants commented on their preferences of color for the button task, but the responses were highly varied as 5 preferred bright colors while 5 preferred lighter and darker colors. There was also high variation in preferences in button size, as 2 preferred a larger size, 4 a medium size, and 1 a smaller size. Further, 8 participants preferred a border, but three did not like the border at all or in some instances. There were also more descriptive words applied to the buttons the participants preferred, such as "muted and professional" (P11), "sleekness and modern look" (P14), and "not scam buttons" (P9). 4 participants made amendments to the qualities of the buttons they preferred after reviewing the designs.

For the text task, 10 participants commented that they preferred a larger font size, 12 participants commented that they preferred darker text, and 11 preferred smaller text spacing. Some characteristics participants preferred of texts were that they were "readable" (P6, P12), "not melted" (P16), and "regular" (P2, P3). 3 participants made amendments to the qualities of the buttons they preferred after reviewing the designs.

3.1.7 Discussion. Overall, relative design acquisition can capture individual visual preferences in a visual context-free scenario that is not dependent on the time of rating of the user. In addition, we see from the high effect sizes that the comparative ratings can be controlled by the design parameter γ (the quality), and that the amount of data might not significantly impact the quality of relative designs generated. We observed more variety and subjectivity in what users preferred for the button task, as opposed to the text task where the majority preferred larger, darker, and appropriately spaced texts. We emphasize that our analysis for the effect on the time of rating (before vs. after) on the comparative rating is preliminary and further experiments are required to validate our approach on the consistency of the comparative ratings when sampled several times for the same user.

3.2 Study 2: Interface Design in a Visual Context-Present Setting

3.2.1 Goals. Study 1 showed that in a visual context-free scenario, relative design acquisition is able to synthesize relative designs for individual users based on their preferences. In Study 2, we want to explore if the results would change if the interfaces were presented in a visual context-present scenario of a shopping website. In addition, Study 1 sampled the data of an individual user using Bayesian Optimization. We are also interested in seeing the effect of the sampling method (Bayesian Optimization vs. random sampling) on relative designs created using a similar experimental protocol. Specifically, we wanted to address the following questions:

(1) Can relative design acquisition capture individual visual preferences of users and be controlled by a design parameter γ (the quality) in a scenario where there is a visual context?
(2) Does the sampling method affect the efficacy of relative design acquisition?

3.2.2 Participants. To address the questions raised, we conducted a study with an additional 16 participants all of which did not participate in Study 1. We recruited 13 males and 3 females with an average age of 20.9 ($sd = 0.9$) within our institution using opportunity sampling. None of the participants had any visual impairments.

3.2.3 Task. In the first task, the website and the button designs were in a blue color scheme, as shown in Figure 7a for the absolute rating task and Figure 7c for the comparative rating task. The design parameters for this task were as follows:

- Saturation of the button color – [0.7, 1.0]
- Value of the button color – [0.7, 1.0]
- Line width of the button border in black – [0.0, 1.5]

Here, the hue of the button color was kept constant at 0.572 for the blue hue. The saturation and value were kept in these ranges for an appropriate blue color that was not too dark. The line width range was selected so that it was appropriately thick for a button of that size on a web-page. In addition, the text "Option" inside the button was kept as black with the same font (Serif) and size (15) so the two buttons in the comparative task are semantically the same.

In the second task, the website and the button designs were in a green color scheme, as shown in Figure 7b for the absolute rating task and Figure 7d for the comparative rating task. The design parameters for this task were as follows:

- Saturation of the button border color – [0.0, 1.0]
- Line width of the button border – [0.1, 3.0]
- Transparency of the text in the button – [0.3, 1.0]
- Transparency of the button background – [0.0, 0.2]

Here, the hue and value of the border color were kept constant at 0.361 and 0.7 respectively and the background color was kept constant at the hue, saturation, and value of 0.222, 1.00, and 0.68 respectively. The line width range was selected so that it ranged from thin to thick; the range of the text transparency was picked so that it was always visible; and the transparency range of the button background in green was chosen so that it remained light. The text "Option" inside the button was kept at constant font (Serif) and size (15).

Specifically, the ranges for the above parameters in both tasks were constrained as our primary consideration is a more practical design scenario where there would be constraints based on website layout and the color scheme. This would restrict for instance the sizes of the buttons as well as the hues of the buttons.

3.2.4 Procedure. Study 2 had $2 \times 2 \times 2$ conditions: 1) the blue and green button tasks; 2) the sampling method based on Bayesian Optimization or random sampling; and 3) $\gamma = 0.6$ or $\gamma = 0.9$. The first two conditions were balanced equally across the 16 participants. The relative designs generated through the two gammas were randomly presented to the participant during each task. Participants were given the prompt of judging the button based on the quality of whether it would attract the most customers within a context of a shopping website. Each condition involved two phases separated by a 3 minute break - 1) the sampling phase that involved only absolute rating tasks; and 2) the comparative rating phase which only involved comparative rating tasks. We started each condition with a tutorial detailing the task followed by 10 absolute ratings and 1 comparative rating to calibrate the user's preferences to different designs.

For each condition, the task was initiated with the sampling phase involving absolute ratings of button designs. For Bayesian Optimization, the task was initialized with 5 random samples and followed by 50 iterations of absolute ratings on designs generated iteratively. For random sampling, there were 55 iterations of absolute ratings based on randomly generated button designs.

After the absolute ratings, participants were asked to take a 3-minute break. Then we generated relative designs and presented them to the participant in a randomized order during the comparative rating phase. For each sampling method, we divided the 50 designs with the absolute ratings into 5 groups of 10 designs each (note the initial 5 random samples were excluded). Then for each group, we took the 3 best designs in absolute rating, resulting in $3 \times 5 = 15$ reference designs for each task. For each reference design, relative designs were generated using all the data collected during the first absolute rating phase with the quality design parameter set to $\gamma = 0.6$ and $\gamma = 0.9$. Then, these comparisons were randomly shuffled and presented to the participant, resulting in $2 \times 15 = 30$ comparative rating tasks in total. The study finished with a short post-study questionnaire, which used the same protocol as in Study 1. Figure 8 illustrates the procedure of this study. Each participant was compensated £10 for their participation and the whole session lasted approximately 45 minutes.

3.2.5 Quantitative Results. We first analyze the distribution of comparative ratings for the designs for each task and sampling method. Table 2 shows the means and standard deviations for each of the conditions of sampling method and gammas for both the blue and green button tasks. Figures 9a and b show the box and whiskers plots for the two tasks. A repeated measures two-way ANOVA was performed with the within-subjects fixed factors of the design parameter γ (the quality) and sampling method (Bayesian Optimization or random sampling) for both the blue and green button tasks with the dependent variable being the users' comparative ratings. None of the analyses violated sphericity by Mauchly's test. For the blue button task, γ was statistically significant ($F_{1,15} = 99.18, p < 0.0001, \eta_P^2 = 0.869$) and the sampling method was statistically significant ($F_{1,15} = 7.569, p = 0.0149, \eta_P^2 = 0.335$) in determining the comparative rating. For the green button task, γ was statistically significant ($F_{1,15} = 156.16, p < 0.0001, \eta_P^2 = 0.912$) and the sampling method was statistically significant ($F_{1,15} = 9.336, p = 0.000801, \eta_P^2 = 0.384$) in determining the comparative rating. For both tasks, the interaction of γ and sampling method was insignificant. Therefore, this analysis suggests that γ is effective at controlling the quality of the relative design in the scenario with a visual context of a website. However, random sampling yielded less strongly preferred comparisons towards the reference design in this experiment.

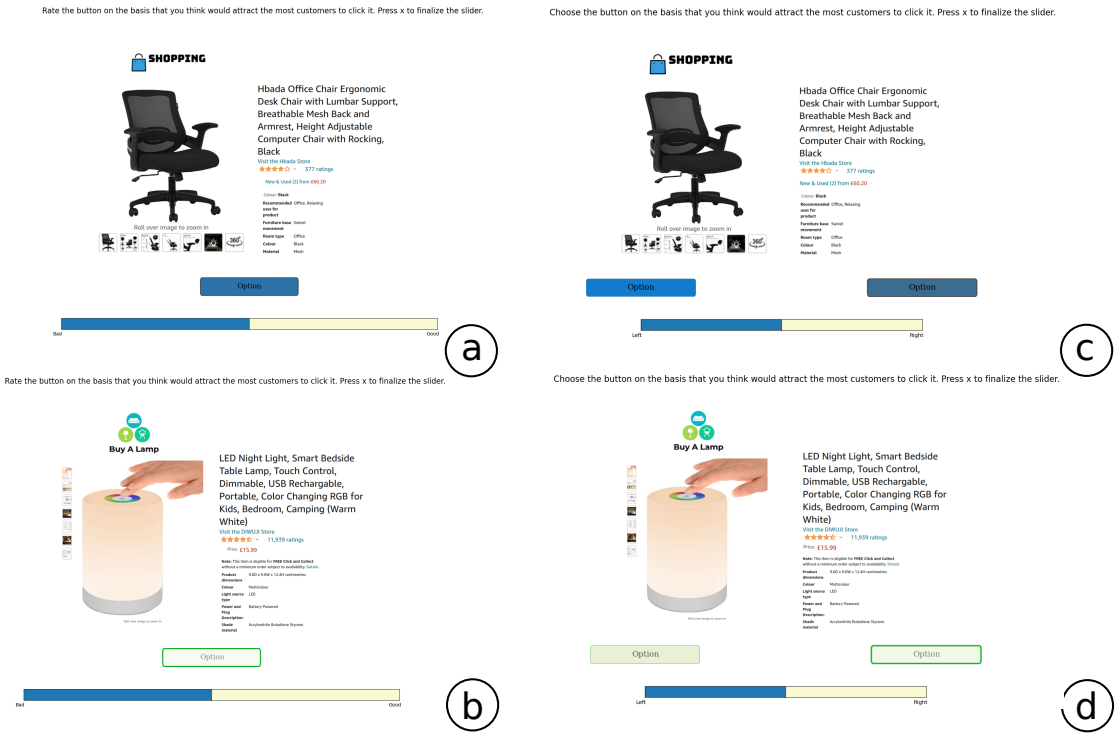

Figure 7: (a) Absolute rating interface for the blue button task; (b) absolute rating interface for the green button task; (c) comparative rating interface for the blue button task; (d) comparative rating interface for the green button task.

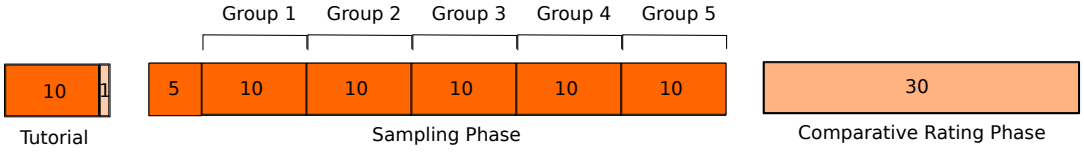

Figure 8: The procedure of Study 2 for each of the four conditions, showing the tutorial, the sampling phase, and the comparative rating phase. The darker orange color represents an absolute rating and the lighter orange color represents a comparative rating. The numbers within the boxes show the number of trials performed.

Table 2: Mean and standard variations for the blue and green tasks for the two γ and sampling conditions.

	$\gamma = 0.6$				$\gamma = 0.9$			
	BO		Random		BO		Random	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Blue	2.649	1.286	3.290	1.658	4.111	1.629	4.437	1.883
Green	3.031	1.365	3.747	1.824	4.398	1.753	5.298	1.849

To analyze why random sampling did not work as well in the comparative ratings we examined the relationship between the reference design absolute rating and the comparative ratings based on that reference design. We used Pearson's correlation to identify any significant correlation between the absolute rating of the reference design and the resulting comparative rating for each of the 2×2 conditions of γ and the sampling method for each of the blue and green button tasks. Table 3 shows the correlations and p-values. The analysis reveals that for $\gamma = 0.6$, there is a statistically significant negative correlation in the absolute rating of the reference design and the corresponding comparative rating. This suggests that when the absolute rating is higher for the reference design, there is more of a tendency for individuals to pick the reference design over the relative one for $\gamma = 0.6$. This postulates that participants were more

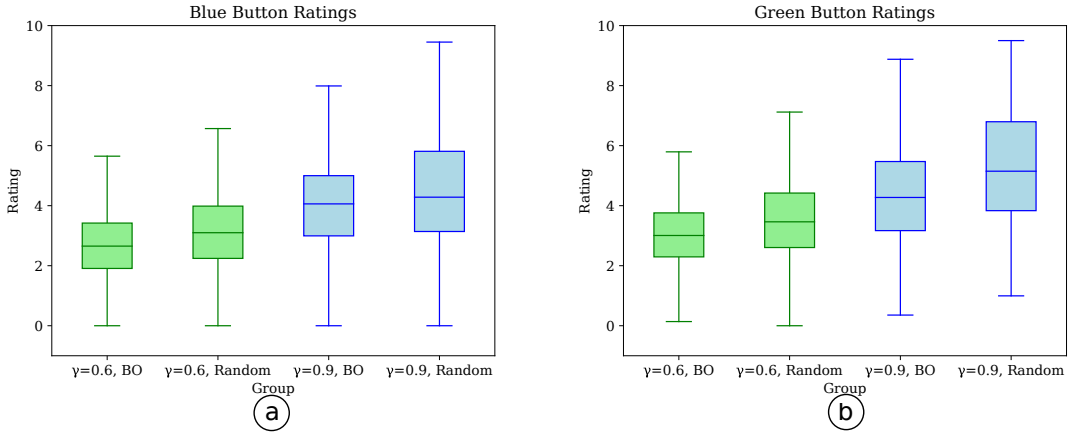

Figure 9: Comparative rating box-and-whiskers plot for the (a) blue button task and (b) the green button task for each condition. The whiskers representing the first or third quartile ±1.5 IQR.

Table 3: Correlation and p-values from the Pearson's Correlation Test for each of the 2×2 conditions of γ and sampling method for each of the blue and green button tasks.

	$\gamma = 0.6$				$\gamma = 0.9$			
	BO		Random		BO		Random	
	ρ	p	ρ	p	ρ	p	ρ	p
Blue	-0.394	< 0.0001	-0.318	< 0.0001	-0.102	0.116	-0.055	0.398
Green	-0.314	< 0.0001	-0.350	< 0.0001	-0.077	0.237	-0.112	0.082

able to discern their preferences on relative design quality to a more optimally attractive reference design as opposed to one that is not. Further, this implies that relative design acquisition would be less effective on reference designs that are not optimal with respect to the design objective in question.

3.2.6 Questionnaire Results. For the blue button task, 14 participants mentioned a combination of light blue colors for the readability of the text, and 11 participants mentioned the thickness of the border, with the preference being not too thick. For the green button task, 14 participants mentioned the readability of the text with a darker text color, and 9 participants mentioned a light green filling was preferable. For the border, 9 participants preferred a thin to medium border, and only 1 participant disliked the border. Only 1 participant (P6) explicitly mentioned using the context to match the colors of the chair and the button in the blue button task. For the blue button task, 5 participants added amendments after reviewing the designs, and for the green button task, 7 participants added amendments.

3.2.7 Discussion. Overall, this study shows that both sampling methods are effective in generating relative interface designs, even in the visual context of a website. There is a large effect size for the comparative ratings for the two γ conditions for both the blue and green button tasks. Correlation analyses suggest that relative design acquisition performs worse when comparisons are carried out with non-optimal visual reference designs. This indicates a guideline in that relative design acquisition may not work as well when the reference design is not a good approximation of the optimum in visual interface applications. In the questionnaire, it was found that there was a good amount of consensus in picking lighter colors for text readability with medium borders in the blue button task and a light filling with a medium border and darker text in the green button task.

3.3 Study 3: Relative Designs for a Group of Users

3.3.1 Goals. Both Study 1 and Study 2 showed that relative design acquisition was effective in generating relative designs in both the visual context-free and visual context-present scenarios. However, each of the studies focused on relative designs generated for each user. In Study 3 we want to examine if our proposed method can create controllable relative designs for a more general population of users. In addition, we are curious to see how the γ controlling the quality of designs impacts the decision time of individuals in deciding which design is more visually appealing. Specifically, we want to address the following questions:

(1) Can relative design acquisition create relative designs that capture the visual preferences of a wider group and be controlled by γ in a variety of tasks?
(2) How does γ affect the decision of individuals to select which design is more visually appealing for different tasks?

3.3.2 Participants. To address the questions raised, we conducted a study with an additional 24 participants—all of which did not

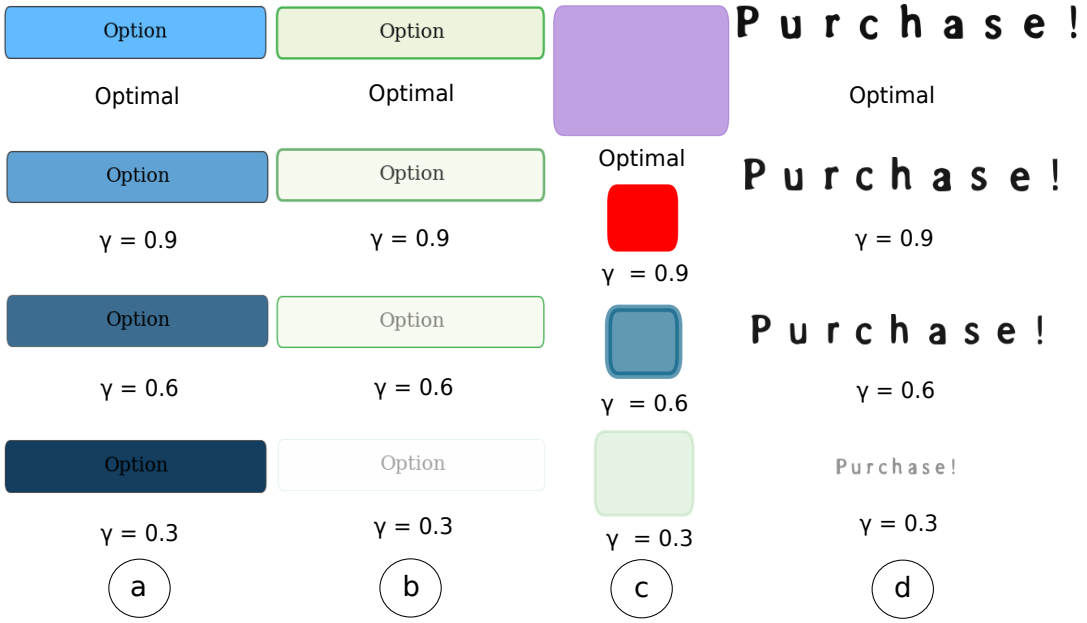

Figure 10: Aggregate blue button (a), and green button designs (b), context-free button (c), and context-free text (d) designs generated, with the reference design and the relative designs for $\gamma = 0.9, 0.6, 0.3$. Note here that the reference designs were generated as near-optimal designs as described in Section 3.3.3.

participate in Study 1 or Study 2. We recruited 16 males and 8 females with an average age of 22.3 ($sd = 2.0$) using opportunity sampling within our institution. There was one participant with an unspecified visual impairment.

3.3.3 Task. Each subject had four tasks to perform: 1) the blue button task with a shopping website context; 2) the green button task with a shopping website context; 3) the context-free button task; and 4) the context-free text task. In each task participants compare two designs side by side. To generate the reference and relative designs for each of the tasks, we first aggregate all the data we have collected in Study 1 and Study 2 for each of the tasks to create a GP for the population of users. Specifically, we used the data collected for the $\gamma = 0.6$ scenarios for both the context-free button and text tasks, and the data collected from the random sampling scenarios for both the green and blue button tasks. These datasets were chosen because they provide the most distinct optima for the four tasks. The reference and relative designs from the aggregated data are shown in Figure 10.

To generate the reference designs which should be near-optimal, we maximized over the posterior means of the Gaussian Process created from the aggregate data. For the context-free button task, we generated 5 distinct optima using the L-BFGS-B optimization algorithm to maximize the posterior mean, and retrieved the posterior mean ratings for each of the distinct optima. However, for the other three tasks, there was only 1 distinct optimum found. To generate 5 visually semi-optimal designs, we applied uniform noise in range $(0, 0.2)$ to each of the design parameters, and used the Gaussian Process to obtain the posterior mean ratings. Using each of the 5 semi-optimal reference designs for each task and the posterior mean ratings, we generated 3 relative designs for each of $\gamma \in \{0.3, 0.6, 0.9\}$, yielding for each task $5 \times 3 \times 3 = 45$ comparisons. The interfaces, design parameters, and prompts were the same in this study as the previous two studies.

3.3.4 Procedure. To counterbalance between the four different task conditions, the two tasks that are visual context-present were paired, and the other two visual context-free tasks were also paired. Each task started with a tutorial with 10 randomized comparative ratings of the interface designs of which the participant was asked to rate which option they preferred more. The decision time was also recorded for each comparison, where the decision time was defined as when the task first started until when the participant had finalized the slider. This was followed by a task of 45 iterations of comparative ratings. The four tasks were equally counterbalanced for all 16 participants. The study finished with a short post-study questionnaire, which used the same protocol as in Study 1 and Study 2, except it now included all four design scenarios. Each participant was compensated £10 for their participation and the entire study took around 30 minutes per participant.

3.3.5 Quantitative Results. We first analyze the comparison ratings for each of the tasks. The box-and-whiskers plots for the four tasks are shown in Figures 11a to d respectively. A repeated measures one-way ANOVA was performed with the within-subjects factor of γ for each of the tasks with the comparative rating as the dependent variable. Sphericity was assessed with Mauchly's test, and if violated, the Greenhouse-Geiser correction was applied. For the blue button, green button, and text tasks, there was statistical significance for γ ($F_{1.51,34.6} = 39.93, p < 0.0001, \eta_p^2 = 0.634$; $F_{2,46} = 47.75, p < 0.0001, \eta_p^2 = 0.621$; $F_{1.15,26.5} = 12.32, p = 0.0011, \eta_p^2 = 0.349$ respectively). We see that for the visual context-present tasks, γ more

Table 4: Mean and standard variations of the comparative ratings for the four tasks and the three γ conditions.

	$\gamma = 0.3$		$\gamma = 0.6$		$\gamma = 0.9$	
	Mean	SD	Mean	SD	Mean	SD
Blue	1.793	1.620	2.598	1.655	3.979	2.003
Green	0.902	1.279	2.223	1.957	3.722	2.277
Button	4.162	2.490	4.327	2.442	4.549	2.373
Text	2.724	2.715	4.051	1.629	4.739	1.619

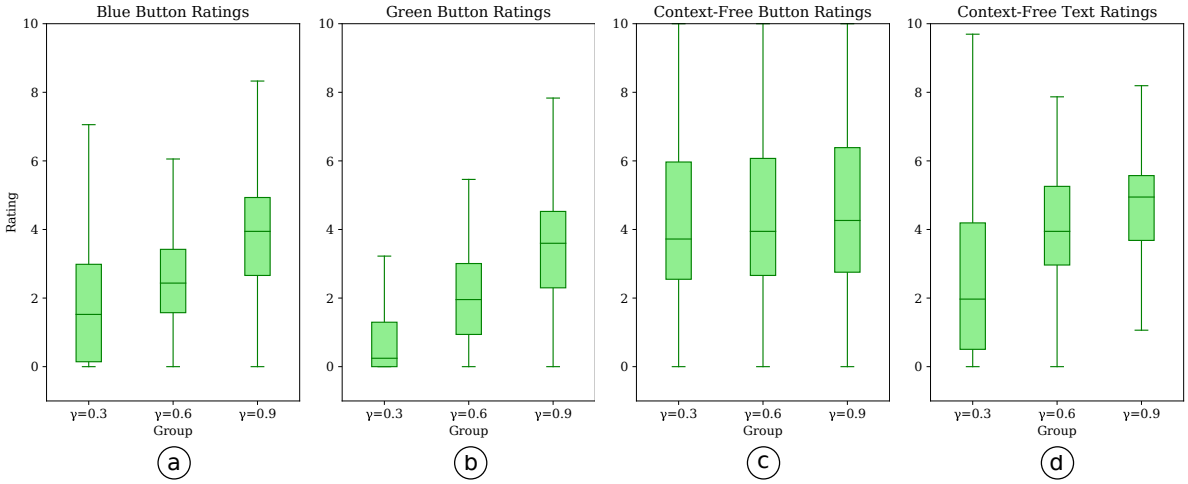

Figure 11: Comparative rating box-and-whiskers plot for (a) the blue button task, (b) the green button task, (c) the context-free button task, and (d) the context-free text task.

Table 5: Mean and standard variations of the decision times in seconds for the four tasks and the three γ conditions.

	$\gamma = 0.3$		$\gamma = 0.6$		$\gamma = 0.9$	
	Mean	SD	Mean	SD	Mean	SD
Blue	3.014	1.666	3.331	2.002	3.697	2.519
Green	2.886	1.329	3.130	2.396	3.586	2.189
Button	4.249	3.085	4.397	3.356	4.522	3.488
Text	3.344	2.188	4.151	2.840	4.376	3.118

effectively controls the comparison ratings. As γ increases, the comparative rating increases so that people are less able to tell which interface design was the visually optimal one (the reference design). For the visual context-free text task, γ is able to control the comparison ratings, but as in Figure 11c, the variance for $\gamma = 0.3$ covers the entire rating range, giving rise to great variability in the preferences for a new group of users. In addition, for the context-free button task, γ has no statistical significance ($F_{1.55,35.7} = 1.661, p = 0.208, \eta_P^2 = 0.0674$).

For decision times, the means and standard deviations for the decision times of the four tasks are shown in Table 5. A repeated measures one-way ANOVA was performed with the within-subjects factor of γ and dependent variable of decision time. Sphericity was assessed with Mauchly's test, and if violated, the Greenhouse-Geiser correction was applied. There was statistical significance in the blue button, green button, and context-free text tasks ($F_{1.57,36.2} = 7.618, p = 0.00337, \eta_P^2 = 0.249$; $F_{1.30,29.9} = 10.16, p = 0.00169, \eta_P^2 = 0.306$; $F_{1.55,35.6} = 5.252, p = 0.0156, \eta_P^2 = 0.186$ respectively). However, there was no statistical significance in the context-free button task ($F_{2,46} = 0.914, p = 0.408, \eta_P^2 = 0.0382$). The high effect sizes and statistical significance suggest that users take more time to make a comparative rating decision as γ increases for tasks with a clear preference consensus. This illustrates that when designs are similar in quality, individuals have more difficulty deciding on their preferences and take longer to make a decision.

3.3.6 Questionnaire Results. For the blue button task, 18 participants commented on the preference for readability of the text with lighter blue colors. 4 participants commented on the preference for brighter blue colors, while 1 preferred a darker color scheme. In addition, 9 participants commented on preferring a border, with 5

preferring medium to thin borders, and 1 preferring a stronger border. 7 participants made amendments after reviewing the designs in this task.

For the green button task, 12 participants explicitly mentioned preferring a darker text for legibility, while only 1 participant stated preferring lighter colored texts. 10 participants also preferred the green filling and 5 mentioned preferring brighter colors for the border. Further, 9 participants made amendments after reviewing the designs in this task.

For the context-free button task, there was a wide range of characteristics that different users preferred. For color, 12 participants reported preferring eye-catching or bright colors, 6 participants preferring lighter and more pastel-like colors, and 2 participants preferring dark buttons. 16 participants commented on preferring borders of various widths, and 5 participants preferred smaller buttons as opposed to 2 participants explicitly mentioning a preference for bigger buttons. There were 9 participants that made amendments after reviewing the designs in this task.

For the context-free text task, 14 participants commented on preferring larger sizes, whereas 5 participants commented liking texts that were not too big or too small. 18 participants said they preferred darker texts and only 1 participant stated a preference for lighter colors. 4 participants commented on preferring texts with smaller spacing. For this task, there were 8 participants that made amendments after reviewing the designs.

3.3.7 Discussion. This study showed that relative design acquisition is able to take accumulated data cross participants to generate relative designs that are effective for a new population of users, most notably better for the two tasks where there is a visual context. It is important to note that this approach works well in terms of the controllability of γ, notably better for visual context-present settings with the large effect sizes. The questionnaire results showed that there was less consensus in the context-free button task, resulting in the lack of statistical significance for γ in the task. In addition, the decision time corresponds closely to γ with the high statistical significance and large effect sizes. When γ is higher for the visual context-present tasks and the text task, it takes longer on average for participants to make a decision. This reflects a result in decision neuroscience where decision time can be used as a proxy for determining the difficulty of a decision task due to the time for discernment between the two choices [15]. Therefore, this study illustrates the ability of relative design acquisition to influence the decision behaviour of a new group of users consistently.

4 DISCUSSION

Study 1 shows that relative design acquisition is able to generate visual relative designs catered to individual user preferences for a visual context-free scenario. In addition, relative design acquisition is able to be controlled by a quality parameter γ, as a γ closer to zero yields comparisons where users are more susceptible to favor the reference design and a γ closer to one yields comparisons where users are less able to discern the reference and relative designs. In this way, the design parameter γ enables control over the quality of the relative design in relation to a reference design. Our exploration of whether relative design acquisition generates relative designs that capture individual preferences consistently across various times of exposure seems promising. However, further investigation is required to confirm the results for longer term exposure.

Study 2 demonstrates that relative design acquisition also extends effectively to a visual context-present scenario in the case of a button design for a shopping website. In this case, γ was also able to control the quality of the relative design generated. Both random sampling and Bayesian optimization yielded effective relative designs that are controllable by γ. However, since random sampling yielded reference designs that were not as optimal (high in absolute rating), this results in higher comparative ratings compared to Bayesian optimization. Study 2 concludes that generating relative designs to a reference design that is not a visual preference optimum may be less effective.

Study 3 shows that by aggregating all the data and creating a surrogate model we are able to find effective relative designs to reference designs representing visual optima for a new group of users for certain tasks. It was not a surprise that the context-free button task was not effective as there is a large variation between individuals of what they found visually attractive, as shown in the post-study questionnaire. There was also a substantial variation of user preferences for the context-free text task. In contrast, for the tasks where the visual shopping website context is provided, the comparative ratings demonstrate a consensus with variation in γ. In addition, γ is able to control the decision times for the blue button, green button, and the context-free text task, where a larger γ yields a greater decision time. This has the design implication that when there is a greater gap in the quality of the two visual designs, users may then be able to choose one choice over the other in less time. Moreover, from a decision neuroscience perspective, γ is able to tune the difficulty of a decision task. It would be interesting to explore this direction in the future with a larger set of participants with greater demographic diversity, perhaps in a crowdsourcing setting in a real-world deployment website scenario.

Overall, relative design acquisition has proven to be effective in generating relative designs not only catered for individual users, but for new groups of users, notably in visual context-present scenarios. However, in the current RDA objective there is no incorporation of variance information in the reference design $\mathbf{x}^*$, as it is only represented by a value of the mean $\mu = \mathbb{E}[f(\mathbf{x}^*)]$ or an estimate of $f(\mathbf{x}^*)$. It would be fruitful to investigate relative design acquisition with an RDA objective that incorporates this variance information, for example, instead of taking the difference in $f(\mathbf{x}^a)$ and its expected quality, a possibility would be to use a quotient form instead:

$$RDA(\mathbf{x}^a|\mathbf{x}^*) = \mathbb{E}_{f(\mathbf{x}^a),f(\mathbf{x}^*)}[|\frac{f(\mathbf{x}^a) - f_l}{f(\mathbf{x}^*) - f_l} - \gamma|]$$

Here, the expectation is taken over both the distributions of $f(\mathbf{x}^a)$ and $f(\mathbf{x}^*)$ in the GP, and therefore incorporates the variance information of both the reference and relative designs.

In terms of applications, it would be interesting to explore relative design acquisition for other interaction and human-in-the-loop tasks where a less subjective design objective function, other than user rating, such as reaction time and attention metrics from eye-tracking, can be assessed. Further, it would be useful to assess relative design acquisition for different visual tasks and assess more sensorimotor reactions of users in the presence of two choices of

similar and different qualities, as controlled by γ. It may also be fruitful to explore relative design acquisition using a deep learning based paradigm, with techniques taking inspiration from interpolation in a latent space.

Although the application with the shopping website may appear limited in scope, the experiments do show that relative designs can be acquired that control user preferences and decision times. We conjecture that relative design acquisition can be applied to different prompts and applications to which it can affect user decisions and choices. However, such investigations are beyond the scope of this paper. Future research directions could include investigating the implications of using relative design acquisition for a real-world task and the resulting user experience, such as exploring whether a generated relative design would affect the proportion of clicks on it versus a reference design. It would also be interesting to understand how a changing visual context could affect the capability of the proposed approach to identify preferences, such as setting up an experiment to investigate how different products on an e-commerce website induce effects on the comparative ratings. Further work could also investigate the fine-grained effects of gamma on user decision making, and hence provide guidance on how small the step sizes for gamma can be set to see a particular comparative rating effect for different applications.

4.1 Ethical Concerns

It is conceivable that relative design acquisition can be used to create dark design patterns, of which the example in Figure 2 can be one such consideration. Dark design patterns are user interface design choices that aim to steer or misdirect users into making unintentional or malicious decisions [3, 20]. For instance, for the blue button and green button tasks, the relative designs generated for $\gamma = 0.3$ and $\gamma = 0.6$ could potentially serve as dark design patterns as they resemble disabled buttons which could result in possible ethical issues [11, 21]. However, although the reference and relative designs exhibited in our studies may exhibit dark design patterns, steering user choices may not always be for malicious purposes. Take for instance widgets being designed relative to one another to prompt the user to click the widget that gives the most helpful option, such as directing users to the most widely used controls in a image editing application. This is an example of relative design acquisition being used to improve users' experience. It is important to highlight the ethical purposes of relative design acquisition and that it should be used to guide users to particular decisions, but not to inherently mislead or misrepresent. It would be an interesting research direction to see if we could inversely infer if a reference and relative design pair would fall into a dark design pattern, for example, by providing a metric indicating if a generated relative design may lead to a malicious purpose.

5 RELATED WORK

5.1 Computational Design in HCI

There has been much work in applying computational approaches and machine learning methods, notably optimization, in human-computer interaction for interface and interaction design tasks. For instance, MenuOptimizer [1] aims to improve user performance by having the designer assisted during the task of combinatorial optimization of menus. Similarly, the tool DesignScape [22] suggests layouts to the designer interactively for position, scale, and alignment of elements. Many of these tools have a human-in-the-loop component where the human designer provides feedback to the design tool, which then updates to propose a new design catering to the tastes of the interacting user. Such tools include Sketchplore [31], a sketching tool which involves real-time design optimization; Forte [6] where topology optimization is used for fabricating shape design; and a tool by Kapoor et al. [14] where classification systems are made more intuitive to users based on user feedback on classification behavior. These tools all have the overarching characteristic that human participation during the optimization process helps improve the quality of generated designs. In particular, the main goal of all the above methods is to leverage both computational methods and human feedback in-the-loop to search for optimal designs that maximize some sort of performance metric. In contrast, we aim to find a relative design to a reference design with a controllable quality factor with respect to some objective.

5.2 Bayesian Optimization in Computational Interaction

Bayesian optimization has been applied extensively in computational design and in human-in-the-loop settings where the objective function is a black-box system (i.e. we do not know the relationship between the user performance and the design parameter beforehand) and expensive to compute, as in the case of user performance to a particular design configuration. Bayesian optimization is a machine learning method that efficiently explores and exploits the design parameter space to find promising new designs based on observations of past design performance. Shahriari et al. [28] provide a detailed review of the applications and the practical aspects of Bayesian optimization. Relevant applications of Bayesian optimization include Brochu et al. [4], which used this approach in a preference gallery scenario to allow users to find the optimal parameters for rendering smoke. Another example is Koyama et al. [16, 17] which used a variant to allow for visual feature optimization in photos with one-dimensional line and two-dimensional planar searches. Finally, Dudley et al. [9] used Bayesian optimization to optimize interface designs in a crowdsourcing setting, and Piovarci et al. [25] employed Bayesian optimization for stylus haptic feedback multi-objective optimization. Overall, Bayesian optimization is an effective statistical computational method for interaction design optimization. However, while we take a statistical computational approach to our problem, we do not have the same aim of finding a design optimum. Instead we focus on finding a relative design to a reference design with a specified quality.

5.3 Interpolation Methods in HCI

The computational approach we take is similar to interpolation methods in machine learning, which have been proposed to find design instances that are a smooth transition from one extreme to the other, such as a rendering a fruit smoothly to morph from an apple to an orange. For instance, in deep learning, there are numerous methods to interpolate over latent spaces, such as variational autoencoders and generative models [2], and applications

to deep feature interpolation in image processing and computer vision [7, 33], as well as in animation [29]. Recently, Ueno et al. [32] proposed a method to create continual and gradual style changes of graphic designs with a generative model, which focuses on the idea of interpolation for deep learning based models. Although our approach is similar to that of interpolation methods, it is different in the sense that our quality parameter controls the performance with respect to some design objective and compares it to a reference design. In addition, our method is more geared to, and robust with, applications where user performance feedback is used, such as in human-in-the-loop settings.

5.4 Perceptual Decision Making

Fundamentally, the purpose of the example to motivate relative design acquisition also presents it as a perceptual decision making problem for humans. In the field of neuroeconomics and decision neuroscience, the effects, factors, and the behavior of decision making have been well researched with respect to sensory inputs being used to create perceptual decisions with discrete categorical variables. Summerfield and Blangero [30] provide a detailed review of the progress in perceptual decision making in the field of decision neuroscience.

In this paper, we focus on the characteristics of decision times when presented with a reference design and a relative design with a specified quality. In the literature, many factors are known to affect decision times. For instance, in a task of reaction-time discrimination of motion direction, elapsed decision time is longer and the decision is less accurate for tasks with less reliable evidence due to the combination of prior information with sensory evidence [12]. In another task of choosing the direction of a noisy display of moving dots, choice certainty was shown to be inversely correlated with reaction time for human participants [15]. Other studies with human and rhesus monkey subjects for visual tasks have shown that urgency affects decision times [26], and that evidence accumulation for decision making has a distinct time-dependent neural firing pattern in the lateral interparietal cortex that affects decision time behavior [8, 27].

6 CONCLUSIONS

This paper has introduced the notion of *relative design acquisition* and formalized it as a computational design problem. Its efficacy in creating visual interface that can steer users' choices was then assessed in three separate user studies. The relative design acquisition method has the advantage that the quality of the relative design to the reference design can be controlled by a quality factor, as captured in the design parameter γ.

The three user studies showed that relative design acquisition is able to create visual relative designs for both visual context-free and visual context-present scenarios catered to individual users, and that it is robust against different sampling methods. In addition, it is effective in generating reference and relative designs controllable by γ for a new set of users, especially in the context-present scenarios. The gap in the quality of designs also affects users' decision time.

Overall, this method has promise in being a general computational design technique and our formalization of relative design acquisition should be widely applicable to a range of interaction design tasks. The results show that an approach based on statistical models, such as Gaussian Processes, is robust and controllable to wide variations of user preferences. Through the three studies, we observe that relative design acquisition has the potential to be deployed in a variety of settings to guide user choices. For instance, less useful tools in a desktop application can be designed to be suboptimal in their visual appearance, and interface designs with more advanced features can be adjusted in quality depending on user proficiency as in training wheels for a user interface [5]. Another possible application would be changing the visual appearance of different paths in a maps application to guide users to choosing alternative routes to alleviate overall traffic in an area. The main advantage of systematically generating relative designs of controllable quality means that visual interfaces can be continuously adjusted according to the context at hand.

REFERENCES

[1] Gilles Bailly, Antti Oulasvirta, Timo Kötzing, and Sabrina Hoppe. 2013. MenuOptimizer: interactive optimization of menu systems. In *Proceedings of the 26th annual ACM symposium on User interface software and technology (UIST '13)*. Association for Computing Machinery, New York, NY, USA, 331–342. https://doi.org/10.1145/2501988.2502024

[2] Piotr Bojanowski, Armand Joulin, David Lopez-Pas, and Arthur Szlam. 2018. Optimizing the Latent Space of Generative Networks. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 600–609. https://proceedings.mlr.press/v80/bojanowski18a.html ISSN: 2640-3498.

[3] Harry Brignull. 2013. Dark Patterns: inside the interfaces designed to trick you. https://www.theverge.com/2013/8/29/4640308/dark-patterns-inside-the-interfaces-designed-to-trick-you

[4] Eric Brochu, Tyson Brochu, and Nando de Freitas. 2010. A Bayesian interactive optimization approach to procedural animation design. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '10)*. Eurographics Association, Goslar, DEU, 103–112.

[5] John M. Carroll and Caroline Carrithers. 1984. Training wheels in a user interface. *Commun. ACM* 27, 8 (Aug. 1984), 800–806. https://doi.org/10.1145/358198.358218

[6] Xiang 'Anthony' Chen, Ye Tao, Guanyun Wang, Runchang Kang, Tovi Grossman, Stelian Coros, and Scott E. Hudson. 2018. Forte: User-Driven Generative Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174070

[7] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. 2019. Homomorphic Latent Space Interpolation for Unpaired Image-To-Image Translation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 2403–2411. https://doi.org/10.1109/CVPR.2019.00251

[8] Anne K. Churchland, Roozbeh Kiani, and Michael N. Shadlen. 2008. Decision-making with multiple alternatives. *Nature Neuroscience* 11, 6 (June 2008), 693–702. https://doi.org/10.1038/nn.2123 Number: 6 Publisher: Nature Publishing Group.

[9] John J. Dudley, Jason T. Jacques, and Per Ola Kristensson. 2019. Crowdsourcing Interface Feature Design with Bayesian Optimization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300482

[10] Joshua Goodman, Gina Venolia, Keith Steury, and Chauncey Parker. 2002. Language modeling for soft keyboards. In *Proceedings of the 7th international conference on Intelligent user interfaces (IUI '02)*. Association for Computing Machinery, New York, NY, USA, 194–195. https://doi.org/10.1145/502716.502753

[11] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174108

[12] Timothy D. Hanks, Mark E. Mazurek, Roozbeh Kiani, Elisabeth Hopp, and Michael N. Shadlen. 2011. Elapsed Decision Time Affects the Weighting of Prior Probability in a Perceptual Decision Task. *Journal of Neuroscience* 31, 17 (April 2011), 6339–6352. https://doi.org/10.1523/JNEUROSCI.5613-10.2011 Publisher: Society for Neuroscience Section: Articles.

[13] Florian Kadner, Yannik Keller, and Constantin Rothkopf. 2021. AdaptiFont: Increasing Individuals Reading Speed with a Generative Font Model and Bayesian Optimization. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3411764.3445140

[14] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1343–1352. https://doi.org/10.1145/1753326.1753529

[15] Roozbeh Kiani, Leah Corthell, and Michael N. Shadlen. 2014. Choice certainty is informed by both evidence and decision time. *Neuron* 84, 6 (Dec. 2014), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

[16] Yuki Koyama, Issei Sato, and Masataka Goto. 2020. Sequential gallery for interactive visual design optimization. *ACM Transactions on Graphics* 39, 4 (July 2020), 88:88:1–88:88:12. https://doi.org/10.1145/3386569.3392444

[17] Yuki Koyama, Issei Sato, Daisuke Sakamoto, and Takeo Igarashi. 2017. Sequential line search for efficient visual design optimization by crowds. *ACM Transactions on Graphics* 36, 4 (July 2017), 48:1–48:11. https://doi.org/10.1145/3072959.3073598

[18] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 147–160. https://doi.org/10.1145/3332165.3347945

[19] Granit Luzhnica and Eduardo Veas. 2019. Optimising Encoding for Vibrotactile Skin Reading. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300465

[20] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 81:1–81:32. https://doi.org/10.1145/3359183

[21] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445610

[22] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2015. DesignScape: Design with Interactive Layout Suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1221–1224. https://doi.org/10.1145/2702123.2702149

[23] Antti Oulasvirta, Niraj Ramesh Dayama, Morteza Shiripour, Maximilian John, and Andreas Karrenbauer. 2020. Combinatorial Optimization of Graphical User Interface Designs. *Proc. IEEE* 108, 3 (March 2020), 434–464. https://doi.org/10.1109/JPROC.2020.2969687 Conference Name: Proceedings of the IEEE.

[24] Antti Oulasvirta, Per Ola Kristensson, Xiaojun Bi, and Andrew Howes. 2018. *Computational interaction*. Oxford University Press.

[25] Michal Piovarči, Danny M. Kaufman, David I. W. Levin, and Piotr Didyk. 2020. Fabrication-in-the-loop co-optimization of surfaces and styli for drawing haptics. *ACM Transactions on Graphics* 39, 4 (Aug. 2020), 116:116:1–116:116:16. https://doi.org/10.1145/3386569.3392467

[26] B. A. J. Reddi and R. H. S. Carpenter. 2000. The influence of urgency on decision time. *Nature Neuroscience* 3, 8 (Aug. 2000), 827–830. https://doi.org/10.1038/77739 Number: 8 Publisher: Nature Publishing Group.

[27] Jamie D. Roitman and Michael N. Shadlen. 2002. Response of Neurons in the Lateral Intraparietal Area during a Combined Visual Discrimination Reaction Time Task. *Journal of Neuroscience* 22, 21 (Nov. 2002), 9475–9489. https://doi.org/10.1523/JNEUROSCI.22-21-09475.2002 Publisher: Society for Neuroscience Section: ARTICLE.

[28] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. 2016. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* 104, 1 (Jan. 2016), 148–175. https://doi.org/10.1109/JPROC.2015.2494218 Conference Name: Proceedings of the IEEE.

[29] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. 2021. Deep Animation Video Interpolation in the Wild. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Nashville, TN, USA, 6583–6591. https://doi.org/10.1109/CVPR46437.2021.00652

[30] C. Summerfield and A. Blangero. 2017. Chapter 12 - Perceptual Decision-Making: What Do We Know, and What Do We Not Know? In *Decision Neuroscience*, Jean-Claude Dreher and Léon Tremblay (Eds.). Academic Press, San Diego, 149–162. https://doi.org/10.1016/B978-0-12-805308-9.00012-9

[31] Kashyap Todi, Daryl Weir, and Antti Oulasvirta. 2016. Sketchplore: Sketch and Explore with a Layout Optimiser. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16)*. Association for Computing Machinery, New York, NY, USA, 543–555. https://doi.org/10.1145/2901790.2901817

[32] Michihiko Ueno and Shin'ichi Satoh. 2021. Continuous and Gradual Style Changes of Graphic Designs with Generative Model. In *26th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, USA, 280–289. https://doi.org/10.1145/3397481.3450666

[33] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. 2017. Deep Feature Interpolation for Image Content Changes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, HI, 6090–6099. https://doi.org/10.1109/CVPR.2017.645

[34] Marc Oliver Wagner, Bernard Yannou, Steffen Kehl, Dominique Feillet, and Jan Eggers. 2003. Ergonomic modelling and optimization of the keyboard arrangement with an ant colony algorithm. *Journal of Engineering Design* 14, 2 (June 2003), 187–208. https://doi.org/10.1080/0954482031000091509

A APPENDIX: DERIVATION OF ANALYTIC FORM OF THE RDA OBJECTIVE

We here detail how to explicitly derive the RDA objective given a Gaussian Process model of f constructed using data $\mathcal{D}$. The only non-deterministic term in the RDA objective is $f(x^a) \sim \mathcal{N}(\mu_a, \sigma_a^2)$, which can be determined using the predictive distribution of the Gaussian process at x^a.

Then, $f(x^a) - (\gamma(\mu - f_l) + f_l) \sim \mathcal{N}(\mu_a - (\gamma(\mu - f_l) + f_l), \sigma_a^2)$. For notational simplicity, denote $K = \mu_a - (\gamma(\mu - f_l) + f_l)$. First we need the following derivation: if $y \sim \mathcal{N}(\mu, \sigma^2)$, then:

$$
\begin{aligned}
F(y) &= \int y \exp(-\frac{(y-\mu)^2}{2\sigma^2}) dy \\
&= \int (x+\mu) \exp(-\frac{x^2}{2\sigma^2}) dx, x = y - \mu \\
&= \int x \exp(-\frac{x^2}{2\sigma^2}) dx + \mu\sqrt{2\pi\sigma^2}\Phi(\frac{x}{\sigma}) \\
&= \int \sigma^2 \exp(-a) da + \mu\sqrt{2\pi\sigma^2}\Phi(\frac{x}{\sigma}), a = \frac{x^2}{2\sigma^2} \\
&= -\sigma^2 \exp(-a) + \mu\sqrt{2\pi\sigma^2}\Phi(\frac{x}{\sigma}) + C \\
&= -\sigma^2 \exp(-\frac{(y-\mu)^2}{2\sigma^2}) + \mu\sqrt{2\pi\sigma^2}\Phi(\frac{y-\mu}{\sigma}) + C
\end{aligned}
$$

$$
\begin{aligned}
\therefore \quad & \mathbb{E}_{f(x^a)\sim\mathcal{N}(\mu_a,\sigma_a^2)}[|f(x^a) - (\gamma(\mu - f_l) + f_l)|] \\
&= \frac{1}{\sqrt{2\pi\sigma_a^2}} \int_{-\infty}^{\infty} |y| \exp(-\frac{(y-K)^2}{2\sigma_a^2}) dy \\
&= \frac{1}{\sqrt{2\pi\sigma_a^2}} (\int_0^{\infty} y \exp(-\frac{(y-K)^2}{2\sigma_a^2}) dy - \int_{-\infty}^{0} y \exp(-\frac{(y-K)^2}{2\sigma_a^2}) dy) \\
&= \frac{1}{\sqrt{2\pi\sigma_a^2}} (F(\infty) - F(0) - F(0) + F(-\infty)) \\
&= \frac{1}{\sqrt{2\pi\sigma_a^2}} (K\sqrt{2\pi\sigma_a^2} + 2\sigma_a^2 \exp(-\frac{K^2}{2\sigma_a^2}) - 2K\sqrt{2\pi\sigma_a^2}\Phi(\frac{-K}{\sigma_a}))
\end{aligned}
$$

Note that here, Φ represents the cumulative distribution function of the Gaussian distribution.