Interaction Design for Reconciling Off-The-Shelf Machine Learning Models with Open-Ended User Needs

March 2024

Hiromu Yakura

Interaction Design for Reconciling Off-The-Shelf Machine Learning Models with Open-Ended User Needs

Graduate School of Science and Technology
Degree Programs in Systems and Information Engineering
University of Tsukuba

March 2024

Hiromu Yakura

Abstract

Machine learning (ML) technologies hold the potential to solve open-ended user needs, thanks to their flexibility that enables adaptive behaviors that cannot be explicitly represented in code and their scalability in replicating the behaviors across multiple computers. However, preparing tailored ML models for each specific need poses challenges, including dependency on substantial computational resources and difficulty in formulating problems to be solved by ML. Therefore, this dissertation introduces the idea of using *off-the-shelf* ML models—models already prepared and accessible publicly—to address open-ended user needs. Nonetheless, off-the-shelf models may not always completely suit user needs. To address this gap, the dissertation proposes an approach that reconciles off-the-shelf models with user needs through interaction design, based on conceptualizing the relationship between humans and ML models as bivariate, not as univariate.

This dissertation presents two strategies as actionable guidance for realizing such interaction designs: 1) utilizing human cognitive processes to identify subproblems tractable with off-the-shelf ML models and 2) fostering human learning to effectively use off-the-shelf ML models. It then showcases concrete examples of interaction designs derived from these strategies against five different domains: video-based lectures, intellectual tasks, photo editing, music composition, and speech transcription. Additionally, by confirming their effectiveness, this dissertation demonstrates that these strategies can derive interaction designs addressing diverse user needs using off-the-shelf models. This work offers flexible and sustainable solutions by giving an eye to possibilities that are outside of existing guidelines for interaction designs with ML, such as those based on human-centered AI. I believe this proposed approach can serve as a *concept* that enables developers and practitioners to generate new interaction designs to address their needs using off-the-shelf ML models.

Acknowledgment

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Masataka Goto, for his invaluable guidance and support. Our association dates back 12 years to when I was a junior high school student participating in the Mitoh program. It was then that I was inspired by his presentation on how computer science research can address influential, real-world issues, such as music understanding. This inspiration led me to seek a research internship under his supervision nine years ago as a first-year undergraduate student. Since then, my learning journey with him started, covering a wide range of skills from how to generate ideas, identify the right research questions, effectively validate hypotheses, and situate my work within the broader research community, to practices on research project and funding management. Despite acknowledging that I was not the easiest PhD student to handle, I am deeply honored to have been his student.

I would also like to express my sincere gratitude to my co-supervisors, Professors Takehito Utsuro and Junichi Hoshino, who have been advising me since my master's course. Their feedback was instrumental in helping me think beyond the boundaries of daily research discussions. My thanks also extend to Professors Buntarou Shizuki and Masaki Matsubara for graciously agreeing to serve on my dissertation committee despite their busy schedules.

Furthermore, I am grateful to all my co-authors, particularly Dr. Yuki Koyama and Riku Arakawa, for our collaboration in the research papers included in this dissertation. My collaboration with Riku, spanning over seven years and resulting in (currently) ten co-first authored papers, has been particularly rewarding. Our friendship, which has allowed us to exchange candid feedback, sometimes down to the minutest details of our sentences, has significantly enhanced both our productivity and my writing skills. Additionally, I extend my thanks to the members of the Media Interaction Research Group at the National Institute of Advanced Industrial Science and Technology (AIST), especially Drs. Masahiro Hamasaki, Tomoyasu Nakano, Jun Kato, and Kento Watanabe. My nine years with the group were filled with stimulating discussions that not only led to several research papers but also fueled my enthusiasm to be a researcher. In particular, discussions with Dr. Jun Kato were pivotal in framing this dissertation within the broader context of HCI research.

I must also acknowledge the financial support that has been crucial to my research. The ACT-X program (JPMJAX200R) by the Japan Science and Technology Agency and the KAKENHI program (JP21J20353) by the Japan Society for the Promotion of Science have been instrumental in my research. I am particularly thankful to Prof. Kenichi Kawarabayashi, director of the ACT-X program, and my advisors in the program, Professors Imari Sato and Yoshihiro Kawahara. Additionally, the Google Ph.D. Fellowship and Microsoft Research Ph.D. Fellowship provided essential support for my research endeavors.

Lastly, but certainly not least, my heartfelt thanks go to my family, whose unwavering support has been the cornerstone of my journey toward a PhD. The journey would not have been possible without the continuous encouragement and support from my parents, for which I am eternally grateful. Thank you very much.

Contents

1 Introduction

> The compensating advantage of a long immaturity is that it enables ontogenetic pathways that incorporate significant amounts of individual learning and cognition, which typically result in more flexible behavioral and cognitive adaptations.
> – Michael Tomasello [300]

Human's brain has incredible plasticity. Our brain continuously changes its neural connections morphologically in response to external environmental and situational changes, enabling us to learn new concepts and acquire new behaviors throughout our lifespan [268]. For instance, it has been observed that London taxi drivers tend to have larger hippocampi, an area associated with spatial navigation, showing a correlation to the length of their careers [190]. Furthermore, there are numerous known cases where individuals, having suffered severe brain trauma resulting in the loss of certain functions, have managed to regain those functions through the reallocation and reprogramming of other brain regions [152, 45]. This special feature of brains allows us to take a sociogenetic strategy. This means that, beyond merely acquiring advantageous traits through genetic mutations, humans have significantly accelerated their evolutionary progress by transmitting and merging knowledge and experiences through social learning [300]. By investing in the brain's plasticity, despite the risks associated with long immaturity——during which time individuals require protection while they learn from the environment—we have developed sophisticated languages and complex social systems [114].

Partially inspired by the mechanisms of the brain that enable this special ability, neural networks were invented. Neural network, one of the machine learning (ML) methods, learns patterns and underlying structures of datasets by updating parameters (often referred to as *weights*), imitating the changes in connection strength between neurons [237]. Its capability to approximate structures representable by any (Borel measurable) function has been theoretically known since the 1980s [121]. However, the concurrent advancements in computing power and the increased availability of large-scale data have significantly broadened the potential applications of the method across various tasks. Consequently, ML technologies including neural networks are now practically employed in various societal contexts, such as credit scoring, local policing, car driving, and online dating [241].

In supporting the expansion of such applications, two key attributes of ML can be highlighted [213]:

Flexibility Through learning from datasets, it is possible to exhibit adaptive behaviors that cannot be explicitly represented in program code. This allows for the development of solutions tailored to complex and dynamic scenarios that pre-coded solutions cannot easily address.

Scalability Once the learning process is complete and the results, such as weights, are saved as models, it becomes feasible to replicate the behaviors in a parallel manner, involving multiple computers. This greatly enhances the capacity to handle large-scale tasks and widespread application deployment.

A symbolic example is demonstrated by a local cucumber farmer in Japan. This farmer has developed its own ML system to classify cucumbers into nine different grades, thereby

reducing the workload of farming tasks [154]. The grading of cucumbers is based on criteria such as thickness and curvature, but these are not quantitatively defined, making it challenging to construct a deterministic classification system even using sensors. The farmer overcame this challenge by collecting images of cucumbers for each grade and using ML to discern these differences. Furthermore, it allows the simultaneous classification of numerous cucumbers and, by replicating the acquired model, enables other farmers to automate their cucumber classification processes. This example suggests the open-ended potential of unseen ML applications not only for tasks in agriculture but also for those in various other fields.

However, there are significant barriers that hinder the further expansion of the applications of ML.

Lack of necessary programming skills Not all practitioners possess the programming capabilities to build ML systems like this farmer.

Limited access to sufficient computational resources ML often requires extensive computation using dedicated hardware, such as graphical processing units (GPUs), which may not be readily accessible to many.

Ambiguity in tasks to apply ML A majority of real-world tasks are carried out based on even more ambiguous criteria than cucumber classification, making it unclear what behaviors should be expected for ML. This ambiguity complicates the construction of appropriate datasets.

Difficulty in comprehensive consideration of environmental factors Even if the expected behaviors for ML are determined, dataset construction involves considering a plethora of environmental factors. For instance, in tasks based on image classification, failing to account for variables such as camera devices and image acquisition conditions in an ecologically valid manner could lead to poorly behaved ML systems.

Considering these barriers, I believe that realizing a world where end-users individually tailor ML models to address their open-ended needs in a grassroots manner remains a challenging prospect.

Instead, in this dissertation, I propose an approach to surmount this challenge by utilizing *off-the-shelf* ML models. Here, off-the-shelf models refer to those that have already been prepared for specific tasks and whose functions are available for access, often via the Internet. By employing these models, my approach seeks to bypass the barriers associated with preparing ML models while addressing the open-ended user needs. A crucial aspect here is *interaction design*, as illustrated in Figure 1. Informed by the insights from human-computer interaction (HCI), the proposed approach bridges the gap between end-users (humans) and ML models (computers) by establishing suitable interactions between them.

1.1 Research scope

The aim of this dissertation is to provide actionable guidance on designing effective interactions using off-the-shelf ML models. As will be reviewed in Section 3.1, there has long been a discourse on the design of interactive systems using ML and artificial intelligent (AI) technologies. For instance, as early as 1962, Engelbart [77] mentioned the importance of

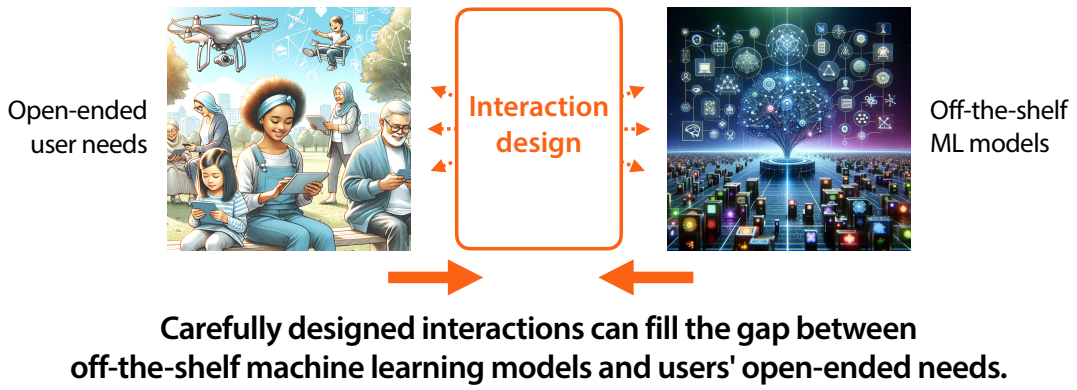


Figure 1: Conceptual illustration of the proposed approach that exploits interaction design to reconcile off-the-shelf ML models with open-ended user needs.

interaction design in the context of proposing the use of computing technologies to augment human intelligence. In the 1990s, the idea of *intelligent user interface* was associated with vigorous discussions on effective interaction design strategies using ML and AI technologies [112, 118]. More recently, a new series of guidelines have been presented considering the contemporary expansion of applications of ML and AI [6, 356]. However, these discussions fundamentally assume the existence of an ML model tailored for a specific task. Therefore, when attempting to circumvent the aforementioned barriers by using off-the-shelf models, they may not be able to be directly applied or may require additional considerations.

In light of this, this dissertation introduces novel strategies for leveraging off-the-shelf ML models to address various user needs. Here, the underlying idea of the strategies is to conceptualize the relationship between humans and ML models not as univariate, but as bivariate. In other words, the previous discussions have mainly focused on the ML side, such as customizing ML models and ameliorating their presentation through interaction design, to make ML models more approachable to humans. Meanwhile, the above conceptualization also considers making the human side more approachable to ML models with the power of interaction design, leveraging the plastic nature of humans. This idea derives the following strategies. The first one is to **utilize human cognitive processes to identify subproblems tractable with off-the-shelf ML models**. This is predicated on the point that within the broad spectrum of open-ended needs, there are often common factors where humans experience difficulty and can benefit from the support of computers. By revisiting human cognitive processes and determining which subproblems, if solved, would lead to effective support for end-users, it becomes possible to exploit off-the-shelf models that address these same factors. Through effective interaction design, these subproblems can then be connected to the ultimate user needs. The second one is to **foster human learning to effectively use off-the-shelf ML models**. This becomes necessary when there is still a gap between the subproblems tractable by the off-the-shelf model and the user needs. In such cases, well-considered interaction design can guide users to enhance their understanding of the behaviors and limitations of off-the-shelf models and skills to cooperate with the models. While this approach might require some effort from the user side, it can be highly beneficial as it fulfills their needs without the burden of preparing new ML models. This perspective is unique to the situations using off-the-shelf models and represents an extension of traditional discussions on interaction

design using ML and AI technologies.

To validate the effectiveness of the two strategies, this dissertation takes the methodology informed by concept-oriented research [282, 119]. Here, *concept* refers to design knowledge that is abstracted from particular instances and has the potential to be appropriated by researchers and practitioners to enable new particulars instantiations [119]. Similarly, this dissertation examines five concrete examples of interaction systems that were designed based on the strategies to clarify their effectiveness and limitations and derive actionable guidance, as outlined in Section 1.2. I believe that by showcasing specific examples in such a wide array of situations and their effectiveness, the utility of the proposed strategies in addressing open-ended user needs is illustrated. Furthermore, this dissertation synthesizes the observations and findings from all these examples to discuss the effectiveness of the proposed approach as well as its limitations and future directions so that it serves as a concept that developers and practitioners can appropriate to deliver a new interactive system harnessing off-the-shelf ML models.

1.2 Dissertation overview

As outlined above, this dissertation explores the effectiveness and limitations of the introduced strategies to present actionable guidance on designing effective interactions using off-the-shelf ML models. While its main body is dedicated to examining the five examples of interaction designs that were generated by the two strategies, the rest of the dissertation is structured in the following manner.

- Chapter 2 first introduces the background of various ML technologies that underpin this dissertation, referencing research I have conducted using tailored ML models. Building on this, it explores how off-the-shelf ML models are created and shared and discusses the difficulties and risks associated with preparing tailored ML models to address specific user needs, highlighting various factors that motivated this research.

- Chapter 3 reviews the research threads on guidelines for interaction design that enable users to benefit from ML technologies and explains how the proposed approach, focusing on off-the-shelf models, can complement these guidelines. It then describes the research methodology of this dissertation and presents the research questions that will be examined in the following chapters.

- Chapter 4 discusses how off-the-shelf models can address the user need to maintain attention in video-based learning situations, especially when learners are temporarily distracted. This chapter demonstrates that an intervention designed based on human cognitive processes enables off-the-shelf models that simply detect moments of distraction to effectively support learners. Additionally, it shows that such an intervention can function without frustrating learners, even if the accuracy of these models is not high. The main body of this chapter is based on a paper presented at the 2021 ACM CHI Conference on Human Factors in Computing Systems [11] in collaboration with Riku Arakawa.

- Chapter 5 explores methods to address the user needs of intellectual workers seeking to avoid task procrastination, based on insights into human cognitive processes related to

persuasive communication. This chapter exhibits that, through carefully constructing interaction designs, it is possible to engage users without preparing tailored models specific to a particular task domain or user characteristics. The main body of this chapter is based on a paper presented at the 2023 ACM CHI Conference on Human Factors in Computing Systems [14] in collaboration with Riku Arakawa and Dr. Masataka Goto.

- Chapter 6 examines how off-the-shelf models can contribute to creativity in photo editing by leveraging human cognitive processes through parametric design. Specifically, it confirms that an interaction design enabled by a new computational framework can bridge the gap between end-to-end ML technologies based on clearly-defined goals and human exploratory processes based on loosely-specified goals. The main body of this chapter is based on a paper presented at the 30th International Joint Conference on Artificial Intelligence [345] in collaboration with Dr. Yuki Koyama and Dr. Masataka Goto.

- Chapter 7 explores interaction designs that facilitate users in learning how to effectively utilize off-the-shelf models. Specifically, the chapter focuses on the context of computer-supported music composition and presents an interaction design that fosters novice users' understanding of musical knowledge and the capabilities of the models. It then confirms that the presented interaction design can promote their utilization of off-the-shelf models, serving as a basis for their creative processes. The main body of this chapter is based on a paper presented at the 24th International Society for Music Information Retrieval Conference [344] in collaboration with Dr. Masataka Goto.

- Chapter 8 applies a similar approach to addressing the user need for efficient speech transcription. It specifically demonstrates how an interaction design that fosters users' skills in respeaking can effectively facilitate collaboration between humans and off-the-shelf models for transcribing unclear speech. The main body of this chapter is based on a paper presented at the 27th ACM International Conference on Intelligent User Interface [13] in collaboration with Riku Arakawa and Dr. Masataka Goto.

- Chapter 9 presents a reflection on the five specific examples and synthesizes their observations and findings. This chapter examines the potency and scope of the proposed approach, elaborating on the promising directions and limitations it presents for generating new interaction designs. It also discusses the ethical considerations that are required to expand the application of the proposed approach.

- Lastly, Chapter 10 concludes the entire dissertation.

1.3 Research contribution

The contributions of this dissertation are three-fold.

Firstly, through five specific examples of interaction designs, this dissertation demonstrates that interaction design can induce the significant applicability of using off-the-shelf ML models. This approach becomes increasingly important in creating a sustainable ecosystem for the application of ML technologies, especially considering the difficulties and risks associated with preparing tailored ML models, as will be discussed in Section 2.3.

Particularly, with the advent of large language and multimodal models, I anticipate that the vast majority of developers and practitioners shift towards using off-the-shelf models via application programming interfaces (APIs) or service as a service (SaaS), rather than preparing and serving their tailored models. This suggests the growing importance of the proposed approach.

Secondly, this dissertation presents actionable strategies for deriving interaction designs to address open-ended user needs using off-the-shelf models. Notably, as mentioned above, previous guidelines on interaction design using ML and AI technologies have primarily assumed the preparation of ML models to meet specific user needs. In contrast, this dissertation illustrates strategies in scenarios where such models cannot be assumed, and thus, different considerations are required. This would be useful not only for researchers but also for developers and practitioners in developing interactive systems tailored to their specific needs.

Thirdly, this dissertation concretely demonstrates how such interactive systems can be implemented, through five actual examples. These systems span different domains such as text, images, and sound and are realized in various implementation forms, including Web applications, browser extensions, and native applications. They can serve as valuable references for the grassroots development of situation-specific interactive systems. Moreover, the effectiveness of these systems is validated using a wide range of methods, offering researchers insights into facets of evaluation for interactive systems employing ML technologies.

2 Background

2.1 Basic paradigm of machine learning

As a fundamental branch of AI, ML technologies encompass several paradigms such as supervised, self-supervised, and unsupervised learning [28, 29]. The most prominent of these is supervised learning. This takes a dataset composed of pairs of input data and corresponding output, discerns patterns or the underlying structure in the dataset, and makes a prediction of the corresponding output for new, unseen inputs. Such a learning process is known as *training*, and the datasets used for this purpose are referred to as training data(sets). Numerous methods have been proposed to realize supervised learning, including support vector machines [28] and gradient boosting [50], which can be utilized for detecting moments when users are distracted from their operation history in PC [347, 346] or predicting appropriate avatar motion from music information [343], for instance. In recent years, neural networks have significantly expanded their application scope, as mentioned in Chapter 1. For example, supervised learning using neural networks can be applied to computer security, specifically, malware detection [349, 350]. In this research, I developed a model capable of malware classification by training a convolutional neural network (CNN), a type of neural network, using image input data derived from malware binary data and output labels indicating the type of malware. Furthermore, by integrating an attention mechanism into the CNN, I showed that it is possible to identify the specific areas in the input data that were crucial in the decision-making process of the model (Figure 2), making the model more interpretable and trustworthy.

Supervised learning can be used to predict more complex outputs, far beyond simple labels, by using Generative Adversarial Networks (GANs) and other generative models [95, 35]. GANs [95] are designed to produce intricate and high-dimensional data, such as images, music, or text, leveraging two neural networks that work in tandem. Large language models are based on a large-scale neural network that is trained on vast datasets of text, learning to predict the next word in a sequence, thereby enabling them to generate text that closely mimics human writing styles [35]. However, the evolution of these methods has led to an exponential increase in the number of parameters within models, which in turn not only necessitates larger datasets for training [68, 231] but also invites various challenges described in Section 2.3.

Self-supervised learning enables training on vast amounts of data without the necessity of labeling them [29], typically by generating pseudo-labels from the input data based on its inherent structure. For example, it can be utilized to train a model capable of distinguishing between various singers' voices with a dataset that consists of a large number of singing voices while lacking explicit information about the singer [351]. In this research, I trained a neural network to consider two singing voice samples taken from the same song as belonging to the same singer, while treating samples obtained by computationally altering in pitch or tempo or extracted from different songs as belonging to different singers (Figure 3). This approach enables the model to learn and differentiate between unique singer characteristics solely based on the properties of the voice samples. Self-supervised learning reduces the burden of preparing output labels, but its dependency on large datasets and substantial computational resources remains unchanged.

Unsupervised learning also operates without labeled data, instead analyzing and finding

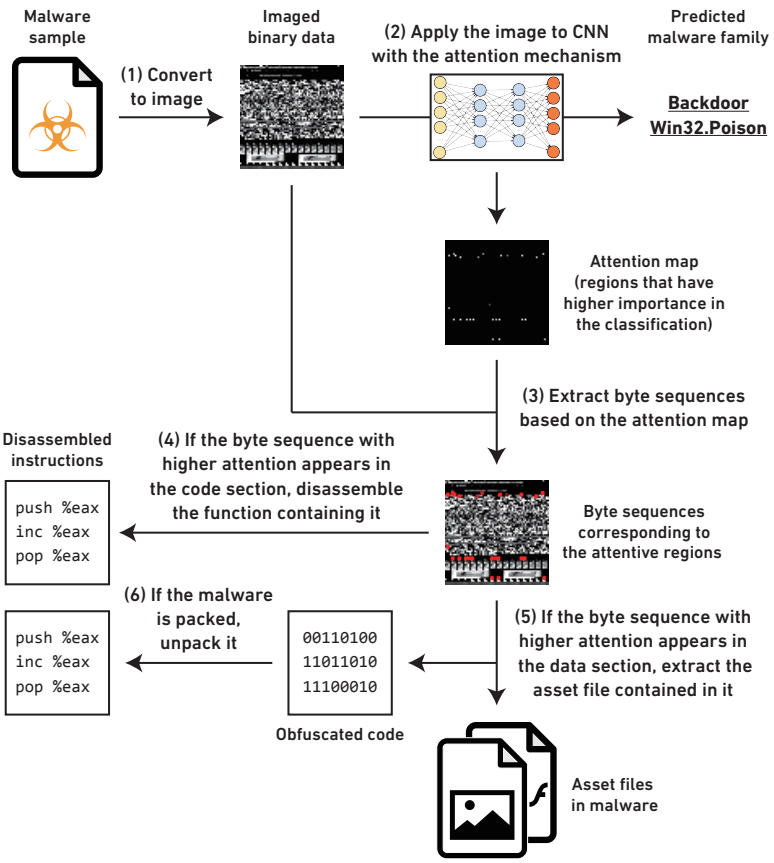


Figure 2: Workflow of classifying malware samples and finding important areas using convolutional neural network [350].

patterns or underlying structures within the data itself, using methods like clustering [28]. It can adaptively process streaming data, continuously learning and updating models based on new, unlabeled information, and thus, often used for anomaly detection [44]. For example, I demonstrated that, in executive coaching, applying unsupervised learning for anomaly detection on clients' behavior information can effectively identify key moments in coaching sessions [12, 10, 9]. This is because anomaly detection can find sudden changes in their behavioral trends, while considering each client's behavioral characteristics, by comparing the current and past behavior information (Figure 4). However, unsupervised learning, especially those operating in an online manner, tends to exhibit behaviors specialized to the data at hand, making them less amenable to scalability—one of the advantages of ML outlined in Chapter 1. Due to this, this dissertation mainly focuses on supervised and self-supervised learning, considering the use of their off-the-shelf models to benefit from ML technologies without preparing tailored models.

2.2 Current landscape of machine learning model creation

As previously mentioned, by saving the parameters acquired through training as a model, we can reuse it, serving as a foundation for one of the key advantages of ML technologies: scalability. Specifically, models trained for particular purposes are referred to as pre-trained models [106]. Among these, those shared in a publicly accessible manner are categorized as off-the-shelf models. The origin of the widespread practice of publicly sharing many pretrained models can be attributed to the fact that programming frameworks and

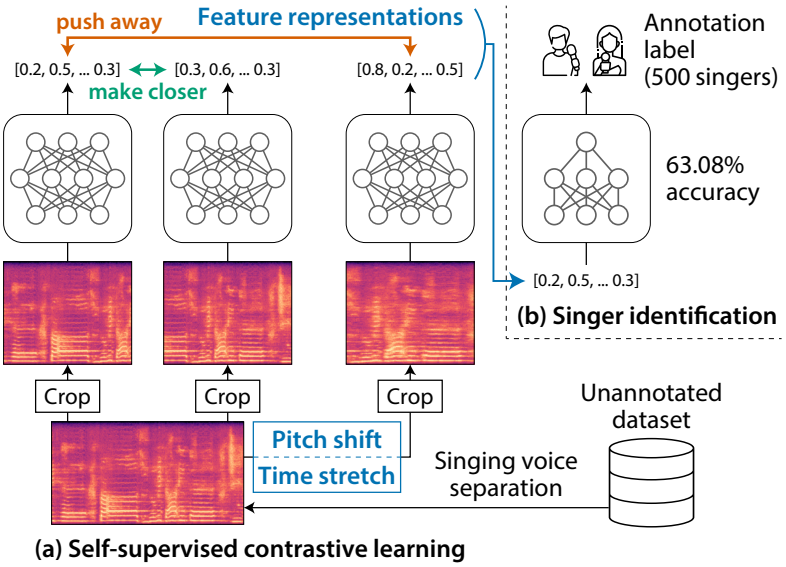


Figure 3: (a) Self-supervised learning enables capturing unique singer characteristics without explicit singer labels. (b) The model can be used for singer identification or other downstream tasks [350].

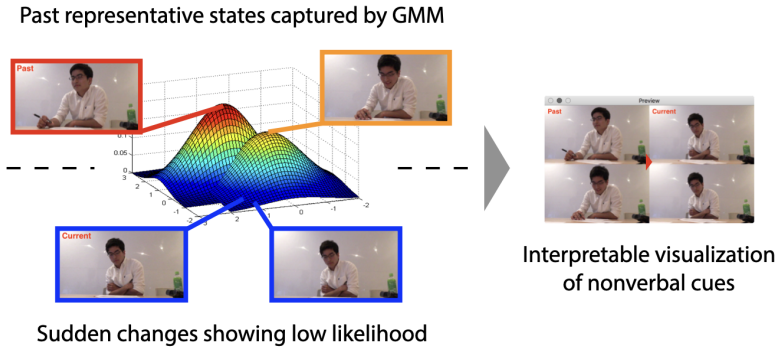


Figure 4: Anomaly detection based on unsupervised learning allows us to capture sudden changes in the nonverbal behavior of a client, which helps coaches infer the client's internal status.

implementations of ML have largely been supported by an open-source community spanning academia, industry, and start-ups [214]. To date, numerous pretrained models have been made available on open-source platforms like GitHub [203], and several dedicated platforms for exchanging off-the-shelf models have also emerged [141]. Among them, the most prominent is HuggingFace, where Castaño *et al.* [43] analyzed more than 380,000 models shared and illustrated a diverse range of off-the-shelf models in domains such as natural language processing, computer vision, audio, and multimodal processing. Meanwhile, a survey by Jiang *et al.* [140] of HuggingFace users indicated that, as the number of parameters in pretrained models increases, it becomes more difficult for users to further customize these models by themselves due to the required computational resource.

On the other hand, functionalities of ML models are also being offered in a manner akin to SaaS, hosted on cloud computing platforms [232, 204]. Known as MLaaS (Machine Learning as a Service), this approach allows users to send inputs via an API, which are then processed by ML models, with the results returned [245, 87]. This trend is partly driven by the increasing costs of computation resources required for the preparation and management of ML models, as detailed in Section 2.3. These services can also be classified under the definition of off-the-shelf models, given their public accessibility. Employing them

enables developers and practitioners to easily use the functionalities of ML models, but it does come with a limitation on the freedom to customize these models. Given the current landscape of creating, sharing, and serving ML models, we can expect that developing approaches to utilize off-the-shelf models against various user needs leads to benefit a wider audience with this ecosystem.

2.3 Difficulties and risks of using tailored machine learning models

While such a wide range of ML technologies has been developed, there remains a significant barrier in preparing tailored ML models to meet each specific user need, as mentioned in Chapter 1. The cost of preparing datasets could be mitigated by self-supervised or unsupervised learning methods, yet access to the computational resources necessary for training is not always readily available. Particularly, it is known that the expressive power of neural networks is constrained in accordance with the exponential increase in the number of parameters [146], and these years have seen the release of models with significantly expanded scales [23]. This also implies an increase in the computational costs of training. For instance, the training cost of PaLM [54], one of the flagship large language models released in 2022, was reported to be around $8 million USD [197], while the training cost of GPT-4 [222], released in 2023, is said to be in the order of $1000 million USD [218]. Furthermore, the increasing computational cost is linked to a rise in carbon dioxide emissions, potentially leading to environmental harm [286, 324]. For example, the training of PaLM [54] resulted in the emission of 271.43 t of CO2. While there are movements to incentivize the development of more environmentally friendly ML models [113, 333], continuing to train ML models in the current manner is criticized for not being sustainable [23, 333]. At the very least, it is impractical for practitioners to prepare tailored ML models for each open-ended user need, considering these environmental and sustainability concerns.

In addition to these issues, another significant barrier is the difficulty in formulating a problem to be solved by ML models to effectively address user needs. For instance, Weiner [325] identifies one of the common pitfalls in the application of AI technologies as *solving the wrong problem*—the attempt to resolve superficial aspects rather than addressing the root cause underlying the needs. Similarly, Joshi [143] highlights the challenges in selecting appropriate solutions or technologies from a holistic perspective. Meanwhile, Nahar [207] points out the challenge in identifying the necessary data for training models, in addition to these points. Dataset construction also frequently serves as a breeding ground for difficulties [254, 289]. For instance, the collected data can often misalign with the actual problem intended to be solved by ML models. This results in *representation bias*, leading to reduced performance or inappropriate behaviors [289, 199]. Furthermore, careful consideration is required to ensure that datasets do not infringe upon others' intellectual property or privacy [227]. In sum, these factors demonstrate the difficulty of engaging in problem formulation and dataset construction by taking all of them into account to prepare tailored ML models.

Even once tailored ML models are acquired, further risks lurk in their deployment, particularly when used in interactive systems accessible to other users. Notably, such systems are susceptible to adversarial examples, which can trigger unintended behaviors by subtly altering inputs with malicious noise [96]. For example, I have shown the possibility

of attacking object recognition or speech recognition models with those that humans feel unsuspicious, as presented in Figure 5, even in real-world scenarios [348, 342]. In such cases, the responsibility of the model creators might come into question. Additionally, techniques have been proposed that allow third parties to extract training data, posing a risk of private data being stolen [85, 39].

Figure 5: Moth-like patches on a STOP sign can make autonomous driving cars misrecognize it as Speed Limit 80 [342].

In conclusion, I argue that preparing tailored ML models to address open-ended user needs is impractical and unsustainable. Instead, considering the availability of many off-the-shelf models as shared resources, as discussed in Section 2.2, it is more reasonable to effectively utilize them. From this perspective, this dissertation considers an approach to reconcile off-the-shelf models with user needs, aiming for a practical and sustainable application of ML technologies.

3 Proposed approach

As introduced in Chapter 1, this dissertation proposes an approach of building effective interaction designs that reconcile off-the-shelf ML models with open-ended user needs. The proposed approach involves two different strategies: *utilize human cognitive processes to identify subproblems tractable with off-the-shelf ML models* and *foster human learning to effectively use off-the-shelf ML models*. This chapter first reviews previous literature about effective interaction design using ML technologies to show how these strategies can complement existing guidelines. On top of the literature, I then formulate research questions to be confirmed through the five examples presented in Chapters 4 to 8 to demonstrate the validity of the strategies.

3.1 Guidelines for interaction design with machine learning technologies

The concept of addressing various tasks through the interaction between humans and intelligent machines was proposed long before the term *machine learning* became widely used. Engelbart [77] referred to this as the H-LAM/T system (Human using Language, Artifacts, Methodology, in which he is Trained), as illustrated in Figure 6. The H-LAM/T system is a framework that encompasses how human intellectual activities can be enhanced by external artifacts, which include not only computers but also other tools like paper and pens. A particularly remarkable aspect of his vision is the emphasis on human *training* as an integral component. He believed that for intellectual activities to be effectively generated through interactions with artifacts, it is also necessary for humans to acquire appropriate skills.

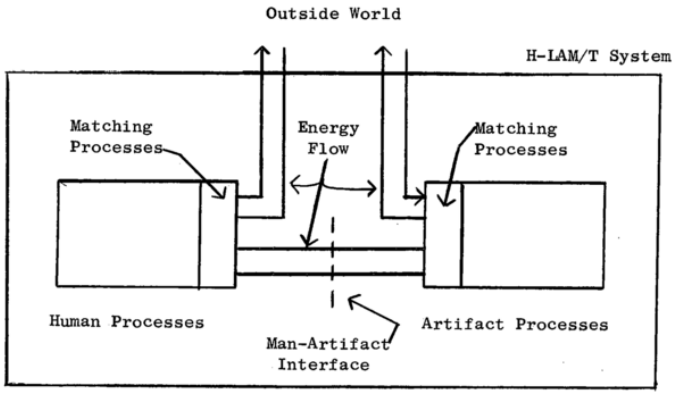


Figure 6: Original illustration of H-LAM/T system (taken from [77])

Later, with the evolution of ML and AI technologies, systems that leverage these technologies to assist users in their tasks have been practically developed, sparking vigorous discussions on interaction design within this context. One prominent trend in this arena was the research on intelligent user interfaces. Hefley and Murray [112] provided the overview of this concept in the proceedings of the first international conference on intelligent user interfaces (ACM IUI). At that time, the primary goal of the collaboration of ML and AI with users was to perform tasks within computers in a simplified and effective manner. This led to a focus on systems that could be used by users without much experience, with many early studies concentrating on systems using conversational agents [137]. For example, Horvitz [122] conceptualized *mixed-initiative user interfaces*, emphasizing the importance

of allowing users to override the decisions of intelligent systems. Additionally, these studies yielded some guidance common to concurrent human-AI interaction, such as being cautious of the iterative refinement of intelligent systems, as it can lead to discrepancies with users' mental models of the systems' behavior [118]. However, one of the challenges at that time was the limitation in accuracy improvement partially due to the lack of large-scale data, which posed constraints on further application expansion [108].

The concurrent advent of ML technologies has completely transformed the landscape. Nowadays, ML is known for exhibiting performance comparable to humans not only in conventional ML tasks like image [109] and speech [338] recognition but also in complex tasks, ranging from Go [273] and race simulations [335] to examinations for medical licensing [93] and professional engineering licensing [238]. This demanded a novel design strategy for human-AI interaction [366], and concurrently, various studies have been conducted to explore effective interaction design. For example, Yang *et al.* [356] identified the common challenges encountered in the design of human-AI interaction, based on interviews with practitioners [355] and a systematic review of past research [354]. Amershi *et al.* [6] have developed a guideline consisting of 18 generally applicable items, and it is officially published by Microsoft. Other industrial companies, like Google and Apple, have also published their guidelines [331], and Yildirim *et al.* [358] examined on how Google's guideline is being utilized by practitioners. They found that the guidelines are not only beneficial for interaction design in production but are also effectively used for internal education and cross-functional alignment. However, it should be noted that these guidelines, like Google's one that includes guidance on dataset construction, mainly assume the self-preparation of tailored ML models. On the other hand, many of the items in these guidelines are still relevant and applicable to interaction design using off-the-shelf models. Given this context, this dissertation aims to complement these guidelines by presenting strategies for effectively utilizing off-the-shelf models and expanding the application of ML technologies in situations where preparing ML models is challenging.

Here, I should mention that the concept of human-centered AI [270, 271] is gaining attention in regard to the overall design of ML systems, which inevitably includes interaction design. This concept advocates for more attention to the benefits and risks of ML systems for end-users and encourages providing controllability to them. Notably, this idea is reflected in the AI strategy of European Union [272], aiming at the construction of intelligent systems that are adaptive to individual users. For the realization of human-centered AI, the application of, for example, explainable AI technologies—which make it possible to provide users with explanations of model behaviors—and privacy-preserving ML algorithms has been expected [191, 339]. On the other hand, the possibility of leveraging off-the-shelf models is not particularly mentioned under this concept, as in the guidelines for human-AI interaction. Rather, the reliance on explainable AI and privacy-preserving algorithms can be understood as the consequence of an implicit assumption that ML models would be customized or served.

I acknowledge that the approach proposed in this dissertation might not be able to be categorized as fully adherent to the definition of human-centered AI, as it focuses not only on making ML models adaptive to humans, but also on adapting human factors to off-the-shelf models. Meanwhile, this approach, which assumes human cooperation and sometimes encourages human learning, could be considered closer to Engelbart's early

idea [77]. Here, some researchers noted that in the 1980s, HCI started to put more focus on technology-oriented research aimed at usability and problem solving, than examining human factors [99, 51]. In this regard, although the proposed approach is aimed at problem solving, it would offer insights into effective human-AI interaction from the attempt to reconnect these two aspects. Specifically, Norman [216] criticized that human-centered design can undermine human adaptability to technology, and Yildirim [358] reported that practitioners using the guideline for human-AI interaction had difficulties reconciling user-centric design process with AI-based solutions. Given that, I hope that this dissertation presents a supplemental perspective to concurrent discussions on human-centered AI.

3.2 From univariate to bivariate understanding of human-AI interaction

To complement the body of knowledge provided by the above literature, this dissertation specifically focuses on interaction design for effectively utilizing off-the-shelf ML models. The underlying philosophy of the aforementioned strategies lies in conceptualizing the relationship between off-the-shelf models and end-users as a bivariate problem. More specifically, the regular approach to designing interaction using ML models has assumed user needs as given or to be clarified in advance, and then optimized ML models and their presentation. In other words, the factors on the user side have been considered as a fixed variable, and the design process is modeled as a univariate problem. In contrast, the proposed approach considers user needs and behaviors to address the needs as adaptable and explores interaction designs that reconcile off-the-shelf models with them. Increasing the degrees of freedom, particularly when one variable—ML models—is constrained to off-the-shelf options, enhances the potential for reaching better solutions. On the other hand, this raises the complexity of the problem, making it desirable to have fair strategies to identify suitable interaction designs. For these reasons, this dissertation proposes two novel strategies with a focus on the adaptability of user needs and behaviors, as illustrated in Figure 7.

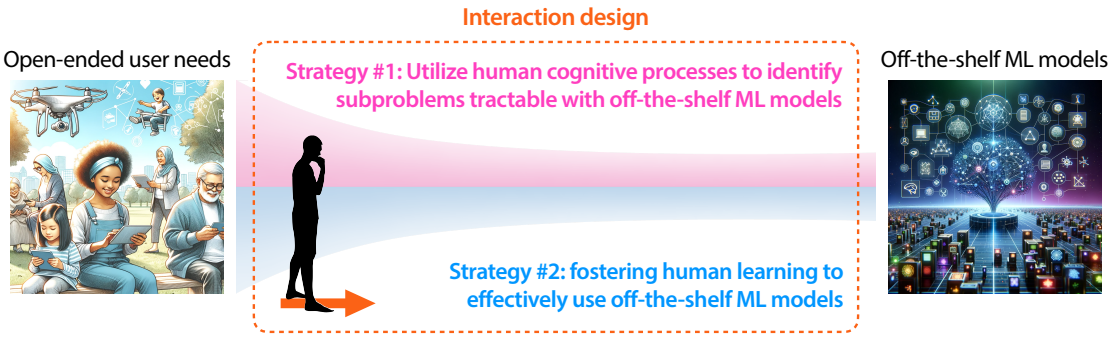


Figure 7: Conceptual illustration of the strategies presented to realize the proposed approach with a focus on the adaptability of user needs and behaviors.

3.2.1 Strategy #1: Utilize human cognitive processes to identify subproblems tractable with off-the-shelf ML models

The first strategy involves understanding and exploiting users' cognitive processes. While I would not assert that the regular approach neglects the users' cognitive processes, the

critical aspect here is the effort to limit the problem that we aim to address with ML models. This strategy can be likened to *matchmaking*, which Bly and Churchill [31] proposed as an approach for interaction design that starts from technological seeds and explores situations where they can be applied. The proposed strategy analogically constrains the problem to be solved within the capability of the off-the-shelf models.

However, it does not necessarily start from the technological side. Instead, understanding users' cognitive processes would precede because it is crucial to identify which subproblems within the user needs should be addressed first. An example can be illustrated by using the concept of *Mindless Computing*, which is actually used in Chapter 4. It was introduced by Adams *et al.* [1] as an alternative to persuasive technologies, being inspired by Kahneman's dual-process theory [145]. More specifically, it exhibited the potential of designing behavior change technologies to function subtly and effortlessly, without requiring users' conscious awareness to be effective, when intervening in their habits or behaviors in domains like health, productivity, or environmental sustainability. By employing this cognitive-process-based approach, we can construct an interaction design that fosters appropriate behavioral change even when no off-the-shelf models can achieve such behavioral change at the desired level. Further, identifying subproblems that can be addressed through interaction design based on cognitive processes, even without the application of machine learning, may lead to uncovering commonalities in the other subproblems that remain unaddressed by interaction design and necessitate the use of ML technologies. Consequently, this strategy increases the likelihood of finding existing off-the-shelf models that have already addressed these subproblems.

3.2.2 Strategy #2: Foster human learning to effectively use off-the-shelf models ML models

The second strategy entails building interaction designs that foster users' understanding of the behaviors and limitations of off-the-shelf models and cultivate their skills for collaborating with the models. Of course, how users comprehend the behavior of computing systems and construct mental models has been a longstanding research topic in HCI [168, 42]. These insights have been leveraged to design interfaces that are easily learnable [234], leading to the recognition of learnability as a significant component of usability [98]. However, interaction design involving the use of ML models presents more complex challenges. This complexity arises because the probabilistic mechanism of ML, which is the source of its flexibility, also complicates the explicit explanation of its holistic behavior. Thus, it is known that users, especially those with non-technical backgrounds, often find it challenging to understand the behavior of interactive systems employing ML [304], which can persist even with the use of explainable AI [76]. Furthermore, while discussions on learnability have traditionally focused on the procedural usage of systems, effectively utilizing off-the-shelf models can demand knowledge and skills that extend beyond mere procedural usages to encompass task-related aspects.

At the same time, the primary motivation for employing off-the-shelf models is to eliminate the burden of preparing tailored ML models. In this context, it would be counter-productive if the learning effort required for end-users becomes substantial. Ideally, derived interaction designs should be functional without human learning, and their effectiveness

should be enhanced as users understand the models' behavior and cultivate appropriate skills. Therefore, this dissertation proposes a strategy that focuses on realizing interaction designs that do not require but rather foster human learning.

3.3 Research methodology

As mentioned above, this dissertation seeks to contribute to the body of knowledge by presenting actionable guidance, rather than explaining the advantages of individual interaction designs. In this respect, there has been a broad discourse in HCI regarding how research on interaction design can contribute to scientific knowledge [41, 329, 17]. For example, *research through design* [78, 376] has been widely employed in HCI to extract knowledge from the iterative attempts to address real-world issues through interaction design. The outcome of such research primarily takes the form of artifacts or systems, sometimes accompanied by insights from actual use cases or demonstration of possible design spaces [89]. Zimmerman *et al.* [377] mentioned that, while research through design can inform knowledge that invites new systems, its connection to such knowledge is not straightforward and unpredictable.

Therefore, I referred to concept-oriented research [282, 119] to inform the research methodology of this dissertation. In particular, the approach of utilizing off-the-shelf ML models by seeing their relationship to end-users as bivariate problems satisfies the definition of *concept*, as it is abstracted beyond particular instances and can be realized in many different ways. At the same time, concept does not assume the universality on its applicability, and thus, clarifying the scope is required. Based on this definition, Höök and Löwgren [119] presented three facets to validate concepts:

Contestable Is the contribution inventive and novel for the academic community in question?

Defensible Is the contribution grounded empirically, analytically, and theoretically?

Substantive Is the contribution relevant to the community in question?

Here, I argue that the contestable and substantive aspects of the proposed approach are presented above by reviewing previous discussions and guidelines for interaction design with ML technologies. In particular, we can see the large new opportunities enabled by specifically focusing on off-the-shelf models, as implied by Sections 2.2 and 2.3. Thus, the following chapters are mainly dedicated to confirming the defensibility of the proposed approach by presenting five concrete examples and abstracting their insights, as outlined in Section 1.2.

3.4 Research questions

The above discussion assumes that the proposed approach can generate new interaction designs that enable the application of off-the-shelf ML models. To ground this in a defensible manner, the effectiveness and limitations of these two strategies need to be examined, which poses the following two research questions.

> **RQ1**: Can the above strategies generate diverse interaction designs against open-ended user needs using off-the-shelf ML models?

	Answer to RQ1	Answer to RQ2
Strategy #1	Presented by deriving interaction designs for video-based lectures (Chapter 4), intellectural tasks (Chapter 5), and photo editing (Chapter 6).	Presented by examining the effectiveness of the left interaction designs in each chapter.
Strategy #2	Presented by deriving interaction designs for music composition (Chapter 7) and speech transcription (Chapter 8).	

Table 1: Overview of the contribution of each chapter to providing answers to the research questions.

Here, the two strategies were introduced to circumvent the barrier of preparing tailored models by utilizing off-the-shelf ML models. Thus, to confirm their capability to generate new interaction designs against the open-ended nature of user needs, it is imperative that the strategies can generate a diverse range of interaction designs. This dissertation attempts to validate the possible scope and limitations by deriving interaction designs for different tasks across multiple domains using the two strategies.

> **RQ2**: Can the interaction designs generated from the strategies effectively address user needs by using off-the-shelf ML models?

Even if we succeed in generating such interaction designs, they must actually address the underlying user needs so that the strategies are actually appropriated by researchers and practitioners for enabling new interaction designs. From this perspective, this dissertation implements these interaction designs as actual interactive systems to confirm their feasibility. Also, based on various evaluation methods informed by HCI literature, I evaluate the effectiveness of each system to answer this research question.

If these two research questions are supported, it confirms the potency of the proposed approach to reconciling off-the-shelf ML models with open-ended user needs through interaction design, as well as the effectiveness of the two strategies introduced to realize this approach. This also means that the proposed approach can serve as a *strong concept* [119], which can contribute to the creation of a world where ML technologies are applied to various tasks by practitioners, as envisioned in Chapter 1. To this end, Chapters 4 to 8 will explore five examples of interaction designs, each derived from the strategies, to address RQ1, and their effectiveness will be confirmed in each chapter, thereby providing answers to RQ2, as summarized in Table 1.

4 Mindless Attractor: An interaction design for helping people maintain attention in video-based learning with off-the-shelf distraction detection models

In this chapter, I explore the application of off-the-shelf ML models in video-based communication. Throughout the past decades, video-based communication has steadily become a viable alternative to face-to-face communication [81, 328]. In particular, schools have leveraged video-based learning to provide educational opportunities for distanced students, as massive open online courses have done [196, 104]. Moreover, the COVID-19 pandemic has precipitated the transition to video-based communication for the purpose of preventing infection [59, 155], especially in the context of education [148, 94]. However, it has been noted that people often have trouble maintaining their attention in video-based communications [160, 161], as they can concurrently perform other tasks, like texting or accessing social media using a smartphone [219]. In particular, it can often occur that people get distracted by side tasks even though they have the motivation to attend in video-based communications [161]. Thus, it would be fruitful if computers could help learners pay attention to a video when they get distracted temporarily as long as they have a basic motivation for taking part in video-based learning from a long-term perspective.

Nevertheless, addressing this need by preparing a tailored ML model, for example, through replicating the behavior of teachers who identify and address inattentive students in a classroom, poses significant challenges. This difficulty arises because such teachers intervene with students based on the combination of their experience and the complex, multimodal information they observe. This amplifies the challenges mentioned in Section 2.3, particularly the serious difficulty in the problem formulation. Then, Strategy #1 presented in Section 3.2 guides us to consider the cognitive process of learners to determine a subproblem that is tractable with off-the-shelf models. Here, looking back to the nature of human communications, we often change the tone of our voices intentionally to draw listeners' attention [208, 291]. Also, humans are known to respond to such paralinguistic cues, which often occur when a topic shift or turn-taking happens [340]. Based on this observation, it is expected that we can help learners return their attention to videos by computationally changing the tone of voice during video-based learning situations. Then, it will help learners pay attention computationally if an off-the-shelf model can simply determine when they are distracted. Fortunately, advances in ML technologies have enabled the automatic estimation of a user's attention level from a facial video [298].

Importantly, this interaction design is also plausible in terms of human cognitive processes related to behavioral change. The power of this design is supported by the concept of *Mindless Computing*—behavior-changing technologies that leverage human biases or unconscious behaviors [1]. Here, I acknowledge that a straightforward approach is to explicitly alert them when they seem not to be paying attention to the video, as Xiao and Wang [337] did. However, such an approach has a strong reliance on user motivation and is likely to fail when learners do not have a strong motivation in video-based learning, especially when the learners' attention is occupied with side tasks. On the other hand, given that Mindless Computing does not consume a user's conscious awareness to be effective, Adams *et al.* [1] stated that this design does not rely on the user's motivation. In addition,

the independence from the user's conscious awareness enables such behavior influencing technologies to work without interfering with the user's primary task, which suits our situation (*i.e.*, use during video-based learning). Furthermore, I argue that such a mindless intervention has a high affinity with sensing modules based on off-the-shelf ML models. That is, if we explicitly alert users, they can be distracted and frustrated by misinformed alerts caused due to erroneous false-positive detection. Here, such behaviors are known to lead users to ignore the result of an ML model [70, 65]. On the other hand, in terms of avoiding the burden explained in Section 2.3, it is challenging to improve the accuracy of the models by customizing them. Still, this mindless intervention can mitigate such negative effects due to false positives because it is based on human cognitive processes and does not necessarily consume users' conscious awareness.

In this chapter, I present a novel interaction design, *Mindless Attractor*, which computationally leverages the nature of our speech communication, and examine its effectiveness in the situation of helping users in video-based learning return their attention to the video. For this purpose, I first conducted an experiment to confirm that the mindless intervention was effective in helping users refocus their attention without consuming conscious awareness. A sensing module using an off-the-shelf ML model was subsequently combined with this mindless intervention to evaluate its effectiveness under the possibility of false-positive detection, compared to a conventional alerting approach. The series of experiments showed the advantages of the presented interaction design, especially in combination with off-the-shelf models. In other words, it was suggested that we can open up the way to apply an off-the-shelf model to address the user need in video-based learning even without customizing the model. I believe that these results present how Strategy #1 in Section 3.2 can inform an effective interaction design, partially answering RQ1 and RQ2.

4.1 Related work

To situate this work, I first examine previous literature on interaction techniques for video-based learning, particularly those focusing on learners' attention. Then, conventional alert-based techniques for drawing human attention are reviewed to discuss why they would not fit our purposes. I also explore previous studies regarding the nature of human speech communication, as this is a foundation of the mindless intervention for drawing users' attention.

4.1.1 Attention-related interaction techniques for video-based learning

As mentioned above, opportunities for video-based communication are increasing, and many interaction techniques have thus been proposed to enhance the experience of such communications. Some prior studies have proposed interaction techniques centering on the context of participants' attention [60, 266, 337], as it has been pointed out that people often have difficulty maintaining their attention during video-based communication [160, 161]. These techniques benefit from the significant effort that has been devoted to estimating participants' attentiveness based on visual cues, such as face movement [298], body postures [364], and gaze [27, 134, 309]. They then use the estimation results to enhance learners' performance, for instance, in video-based learning, given that learners' attention and engagement are strongly related to their learning performance [15, 60].

For example, Gaze Tutor is a gaze-reactive intelligent tutoring system for video-based learning [60]. Using a conventional eye tracker, it estimates the learner's attention level from gaze direction based on a simple rule assuming that off-screen gaze patterns imply distraction. When the system detects that the learner is not focusing on the video, the tutor agent stops the video and alerts them explicitly (*e.g.*, by saying "Please pay attention"). Although their experiment showed its effectiveness in reorienting participants' attention, the intervention method left room for improvement, as the authors mentioned in their discussion. Specifically, they found individual differences in the efficacy of the alert-based intervention, including that some participants never followed the alerts. Accordingly, the authors noted that alternate intervention approaches, including indirect feedback, could be implemented. Another example that computationally utilizes the estimated attention level during video-based learning was provided by Sharma *et al.* [266]. Similar to Gaze Tutor, their system provides users with direct feedback, such as simple red rectangles on the screen, with the purpose of improving users' attention.

As can be inferred from these studies, previous research has mainly considered explicit alerting as an intervention method for video-based learning. However, the findings from these studies complement our concern, which is discussed at the beginning of this chapter based on the results of Xiao *et al.* [337]. That is, such interventions rely on users' motivation and may not work effectively when learners are distracted by side tasks. This necessitates the exploration of a better interaction design to realize a situation where an off-the-shelf ML model that detects a distracted moment can help them.

4.1.2 Alerting techniques for drawing human attention

Drawing users' attention is one of the crucial components of HCI, not being limited to video-based learning. Many researchers have dealt with a wide range of topics in this area, such as Internet advertisements [211], smartphone notifications [283], and alerting systems [103]. Consequently, previous studies have developed many methods suitable for individual situations using diverse perceptual modalities. One of the most popular strategies is to provide users with visual stimulation. For example, Red Alert is a visual alerting system that uses a translucent orange-red flash to mask a screen, designed to warn pilots of potential collisions in air traffic control [258]. Audio stimuli have also been favorably employed as a means to alert users. BBeep is a collision-avoidance system that can emit a beep sound to alert pedestrians around a visually impaired user to clear the way [147]. Another strategy is the use of the tactile modality. BuzzWear is a wrist-worn tactile display to notify users on the go by combining different parameters of the tactile stimulus [167]. As can be observed in these examples, most systems adopt explicit stimuli to notify users, assuming that they will take action after their attention is drawn to the target.

However, Adams *et al.* [1] pointed out that such alerting strategies would not be optimal when used within persuasive technologies designed to influence user behavior. Unlike critical situations (*e.g.*, air traffic control), where it can be expected that users will be motivated to follow an alert from a computer, not all scenarios for inducing behavioral change can assume that users are highly motivated to do so. In such cases, an alert that requires the user's conscious awareness and effort to work would likely fail due to a temporary lack of motivation or potentially counteract positive aspects of the intervention by frustrating them.

Thus, the authors recommended the Mindless Computing strategy of leveraging human biases or unconscious behaviors, which diminishes reliance on users' conscious awareness. It also enables intervening with users without interfering with their primary activities, whereas alerting users explicitly can interrupt such activities. Furthermore, they complimented the advantage of the mindless intervention by mentioning its long-term effectiveness, which persists even though users are aware of the biases behind the interventions [320]. Then, our next question would be how to implement such an intervention in the situation of video-based learning, which leads us to make use of the nature of human speech communication.

4.1.3 Speech communication techniques for drawing human attention

Speech is one of the most natural modalities of human communication. It consists not only of linguistic aspects but also of paralinguistic aspects, such as pitch, volume, and speed, which play an important role in conveying nuance or emotion [301]. While the use of paralinguistic aspects is a natural habit in our daily communication [235], it is also a common practice to intentionally create changes in such paralinguistic parameters during speaking to draw listeners' attention [136]. The relationship between speech parameters and their effects in drawing attention has generated considerable research interest in understanding human speech communication [208, 291]. For example, Xu [340] suggested that an increase in pitch when starting a new topic can draw listeners' attention. Moreover, a similar effect of drawing attention has also been observed in infants hearing their mothers' speech, who naturally vary their pitch [288]. The idea that humans unconsciously respond to paralinguistic cues is further supported by Zatoree and Gandour [365], who verified that human neural mechanisms are sensitive to such spectral and temporal acoustic properties.

Based on these results, I speculate that leveraging this nature of human speech communication by computationally varying speech parameters can naturally draw listeners' attention. More specifically, if a person who is losing their attention to a video hears speech with altered pitch or volume, they will respond to such a change, regardless of their motivation to pay attention. Such an intervention aligns with the concept of Mindless Computing [1] and thus is expected to work when side tasks occupy users' attention. In the following section, I elaborate on the rationale for using human speech alterations to draw attention in video-based learning situations.

4.2 Mindless Attractor

In this chapter, I present *Mindless Attractor* for the purpose of helping users in video-based learning return their attention to the video. Guided by the strategies to utilize off-the-shelf ML models, it leverages the nature of speech communication to intervene with users based on the concept of Mindless Computing [1], This section explains the details of Mindless Attractor, starting by discussing its design rationale and requirements.

4.2.1 Why mindless?

As stated at the beginning of this chapter, the aim of the presented interaction design is to support video-based learning, given the growing demand for it, by establishing a suitable computational intervention for users who are not paying attention to the video. The difficulty is that we cannot assume all users to be highly motivated to follow such an

intervention for maintaining attention when distracted by side tasks, even though they have a basic motivation for video-based learning. Thus, as mentioned in Section 4.1.2, conventional alerting approaches would not be optimal, and we need to consider an intervention approach that does not rely on users' motivations. Here, even when a user is not focusing on the video, interventions that interrupt the user should be avoided since such approaches might lead them to miss subsequent content. These points led us to adopt an interaction design based on Mindless Computing [1] that leverages human biases or unconscious behaviors to induce behavioral change. Since such an intervention does not consume the user's conscious awareness to be effective, it is considered less reliant on their motivation to pay attention. Moreover, it enables us to design a less interruptive intervention than explicit alerts, as Adams *et al.* [1] confirmed that their mindless approach using auditory feedback could influence people's behavior when talking without annoying them. Then, we can implement an interactive system that effectively helps users by utilizing an off-the-shelf model that simply detects their distracted moments.

Furthermore, it can be presumed that the mindless intervention reveals a new advantage when integrated with a sensing module based on off-the-shelf models, as mentioned at the beginning of this chapter. More specifically, although ML technologies enable various sensing scenarios, humans tend to evaluate mistakes of such ML-based systems more severely than human mistakes [70]. In addition, the trust that ML-based systems lose as a result of their failure is usually greater than the trust they gain from their success [361]. Consequently, people often become less hesitant to override outputs from ML-based systems after seeing their failures [65]. Moreover, it has been suggested that people with a high cognitive load will have less trust in interactions with ML-based systems [372]. These discussions imply the risk posed by the false-positive detection of the sensing module in intervening with users—that is, mistakenly alerting them in an explicit manner during video-based learning situations would frustrate them and lead them to disregard the alerts. On the other hand, since the mindless intervention does not consume conscious awareness, unlike the alerting approach, it might mitigate the negative effects caused by false positives.

4.2.2 Designing Mindless Attractor

The above discussions rationalize the use of the mindless intervention to transform complex user needs in video-based learning situations into a subproblem tractable by an off-the-shelf ML model. In this presented interaction design, I specifically leveraged the nature of human speech communication as human biases or unconscious behaviors behind the intervention, based on Section 4.1.3. Then, we need to consider several requirements to make the intervention work while video-based learning, as follows.

Avoid interruption due to interventions. Considering that video-based learning is sometimes delivered in the form of live streams or in a synchronous manner [20], interrupting users due to interventions should be avoided, as it can cause them to miss information and counteract the aim of helping them pay attention. This requirement is one reason to eliminate the use of alerting approaches, as their interruptive aspect was discussed in Section 4.2.1.

Use a modality that users will not neglect. To intervene with users who are not paying attention to the video, it is important to use a modality that is always reachable

to users. In this regard, though it is possible to leverage human perceptual bias to implement the mindless intervention by showing something on a display, this would not be suitable because the user can take their eyes off the display, especially when performing side tasks using a smartphone [219]. On the other hand, it seems more unlikely that the user would not hear the audio due to muting it while in video-based learning situations.

Function without external devices. Though the use of external devices would extend the range of possible interventions, such as using a tactile stimulus [167], it raises an additional cost to utilize the interventions. Therefore, it is desirable to implement an intervention that could be integrated into video-based learning situations without requiring external devices.

As reviewed in Section 4.1.3, it has been suggested that humans unconsciously respond to paralinguistic cues in speech, such as a change in pitch, volume, and speed. For the presented interaction design, the perturbation of the pitch or volume of the voice in the video was used to help users refocus their attention. Here, speed change was not employed because it would be difficult to maintain time-series consistency when video-based learning is conducted in a synchronous manner (*e.g.*, live lectures [20]). In addition, the perturbation is enabled and disabled repeatedly when the user is seemingly not paying attention to the video, as Adams *et al.* [1] emphasized the importance of cues to trigger different perceptions and sensations in designing mindless approaches. Otherwise, if we activated the perturbation once when the user became distracted and kept it thereafter, the user would have less opportunity to refocus their attention as they became acclimated to the changed pitch or volume.

4.2.3 Implementation

Based on the above consideration, I implemented an interactive system that presents the mindless intervention by using Python and PyAudio[1] to perturb the audio signal in real time. The audio signal was captured in 16 kHz, and the perturbation process was activated each $\frac{1}{16}$ sec to ensure that the perturbed signal was delivered without significant delay. The pitch shift was performed using a library named rubberband[2] through time-shifting and resampling the signal via Fourier transform. The volume change was performed by directly multiplying the waveform double or halve. In addition, as mentioned in Section 4.2.1, the mindless intervention is expected to incorporate a sensing module based on an off-the-shelf ML model that monitors users' behavior and detects when they are distracted. The detailed implementation regarding how the sensing module is constructed using an off-the-shelf model is later explained in Section 4.5.3.

4.3 Hypotheses

Up to this point, I have introduced Mindless Attractor, which is designed to intervene with users during video-based learning that incorporates a sensing module based on off-the-shelf ML models. It computationally perturbs the pitch and volume of the voice in the video in

[1] https://people.csail.mit.edu/hubert/pyaudio/docs/
[2] https://github.com/breakfastquay/rubberband

real time to refocus users' attention when they seem distracted from the video. The design rationale for the presented interaction design, which was discussed in Section 4.2.2, imposes the following hypotheses to be verified to ensure its validity and effectiveness.

First, as discussed in Section 4.2.1, this interaction design is based on the concept of Mindless Computing [1] so as to ensure that the intervention works without relying on user high motivation and without interrupting users. To satisfy these points, we should examine whether Mindless Attractor can influence users' behavior in a mindless manner, *i.e.*, without consuming their conscious awareness.

> **H1**: Mindless Attractor is an effective means to refocus the attention of users in video-based learning situations without consuming their conscious awareness.

If H1 holds, we have two choices for inducing a behavioral change in users (*i.e.*, drawing their attention back to the video): alerting users in an explicit manner or intervening in a mindless manner. Here, as discussed in Section 4.2.1, it is expected that the mindless intervention will be favored over alerting approaches when combined with a sensing module based on an off-the-shelf ML model for detecting users' distraction. More specifically, the fact that such a sensing module may produce false positives implies the risk of mistakenly intervening in users, which can be annoying when we alert them explicitly. This point guides the following second hypothesis:

> **H2**: Mindless Attractor is not only an effective means to refocus users' attention but is also preferred by users when combined with a sensing module based on an off-the-shelf ML model, while the alerting approach is not accepted.

If these hypotheses are supported, we can pave the way for intervening with users in real time to support their participation during video-based learning. Also, it becomes an illustrative example showing the possibility of addressing such a user need with an off-the-shelf model by carefully designing an interactive system based on human cognitive processes, as discussed in Section 3.2. With this motivation, I evaluated these hypotheses by conducting a series of experiments.

4.4 Experiment 1: Evaluation of H1

4.4.1 Design

To evaluate H1, I conducted an experiment that replicated video-based learning situations. A within-participant design was used to compare a treatment condition using Mindless Attractor with a control condition that did not intervene in participants. Then, H1 is supported if the following two points are confirmed: Mindless Attractor helps participants refocus their attention, and Mindless Attractor does not consume participants' conscious awareness.

4.4.2 Measure

Two measures were prepared in correspondence to the above two points to be confirmed: recovery time and cognitive workload.

Recovery time This metric indicates the time that it took for participants to return their attention to the video after losing focus. If Mindless Attractor helps participants

refocus their attention, the time that they are distracted should be shortened in comparison to the case in which no intervention was taken. To compute this metric, human annotations were collected for each participant, denoting whether the participant was paying attention or not. As explained later in Section 4.4.5, an experimenter who observed the state of the participants annotated in real time so that the recovery time could be calculated later.

Cognitive Workload This metric was used to evaluate whether Mindless Attractor consumed the participants' conscious awareness or not. Here, measuring cognitive workload is common in the previous studies proposing alerting approaches [258, 167]. Whereas they aimed to show that their proposed approaches exhibited lower workload compared to other possible approaches, this experiment compared the metric between the control and treatment conditions. If the cognitive workload in the treatment condition is not significantly different from that in the control condition, it suggests that Mindless Attractor does not consume participants' conscious awareness. In this experiment, the NASA-TLX questionnaire [107, 37] was used to measure cognitive workload, in the same manner as the previous studies [258, 167].

Note that it would be possible to evaluate whether Mindless Attractor consumes the participants' conscious awareness by asking them whether they noticed the perturbation. However, to do so, it was necessary to conceal from the participants that they would be subject to an intervention, which would create an unrealistic situation if we consider the practical applications of the presented interaction design. More specifically, it is unlikely that users in video-based learning would be subject to interventions without opt-in consent. In other words, they would use Mindless Attractor of their own accord to focus on videos or at least would be notified about the possibility of the intervention. In addition, as mentioned in Section 4.1.2, Adams *et al.* [1] explained that the mindless approaches work regardless of whether a user knows their mechanisms or not, as they do not depend on the user's conscious awareness. Thus, the participants were notified beforehand that they would be subject to interventions and were presented with this measure based on NASA-TLX afterward.

4.4.3 Material

To replicate a video-based learning situation, a video recording of a 30-minute lecture on urban sociology was prepared. As this experiment was conducted remotely, the video was presented to the participants using the screen-sharing function of Zoom.[3] By following the implementation described in Section 4.2.3, a client software was also prepared that modifies Zoom's audio output to perform the intervention. This software captures and perturbs the audio output in real time when it receives an activation command from a control server via WebSocket. Here, a pilot study was conducted in the same manner as Adams *et al.* [1] to find the appropriate parameters for intervening without causing distractions. This led to the implementation of the following four perturbation patterns: halving or doubling the volume and lowering or raising the pitch by one tone. The software then activates one of the four patterns randomly so as to enable the comparison of their effectiveness in helping the participants refocus their attention. Since Zoom automatically removes noises and extracts voices, it was confirmed that a naïve implementation of pitch shifting based on a fast Fourier

[3] https://zoom.us/

transform would be sufficient for the purposes of this experiment. An experimenter console was further prepared in the control server to record annotations concerning whether the participant was paying attention or not. The console was implemented to enable sending the activation and deactivation commands to the client software when the participants started to divert their attention from the video and refocus their attention, respectively.

4.4.4 Participants

This experiment involved 10 participants, three of whom were female. They were recruited via online communication in a local community where over 100 university students gathered. As described later in Section 4.4.5, the current experimental procedure required participants to be observed by a remote experimenter so that their state of attention could be annotated. Therefore, they were asked to prepare a PC with a webcam in a quiet room and to enable their faces to be captured.

4.4.5 Procedure

Each participant underwent one session of watching the 30-minute video using a computer connected over Zoom, as mentioned in Section 4.4.3. To replicate the usual situation of video-based learning, in which learners have some reasons to watch the video, the participants were told in advance that they would be asked to write a few sentences summarizing the video. At the same time, they were asked to bring their smartphones and told that the use of smartphones was not prohibited so that they could be distracted as usual [219]. Then, as depicted in Figure 8, each session was divided into two parts of 15 minutes each: one with no intervention and the other involving interventions. To normalize the order effect, the order of the two parts was balanced: five participants first experienced the part with no intervention, and the others first experienced the part involving interventions. After each part, the participant was asked to write a summary and fill out the questionnaire measuring cognitive workload. Note that these two parts do not correspond to the control and treatment conditions, as explained in the following paragraphs.

In the part involving interventions, an experimenter observed the state of a participant, including their use of smartphones, and annotated whether they were paying attention to the video or not. When the experimenter pressed a button on the experimenter console to record the timestamp at which the participant diverted their attention from the video, the console assigned either the control or treatment condition with a 50% probability of each. Note that the selected condition was concealed from the experimenter in order to avoid the experimenter bias in the annotations. If the treatment condition was assigned, the console sent the activation command to the client, and the client then repeatedly enabled and disabled one of the four perturbation patterns every 3 s, as explained in Section 4.2.2. This intervention continued until the client received the deactivation command indicating that the experimenter pressed another button to record the participant's recovery from the distraction. On the other hand, if the control condition was assigned, no command was sent to the client. Consequently, based on the assigned conditions and the recorded timestamps, the recovery time could be calculated and compared. The other part (with no intervention) was prepared to evaluate the cognitive workload. Its cognitive workload score was compared with that of the part involving interventions, which were activated on

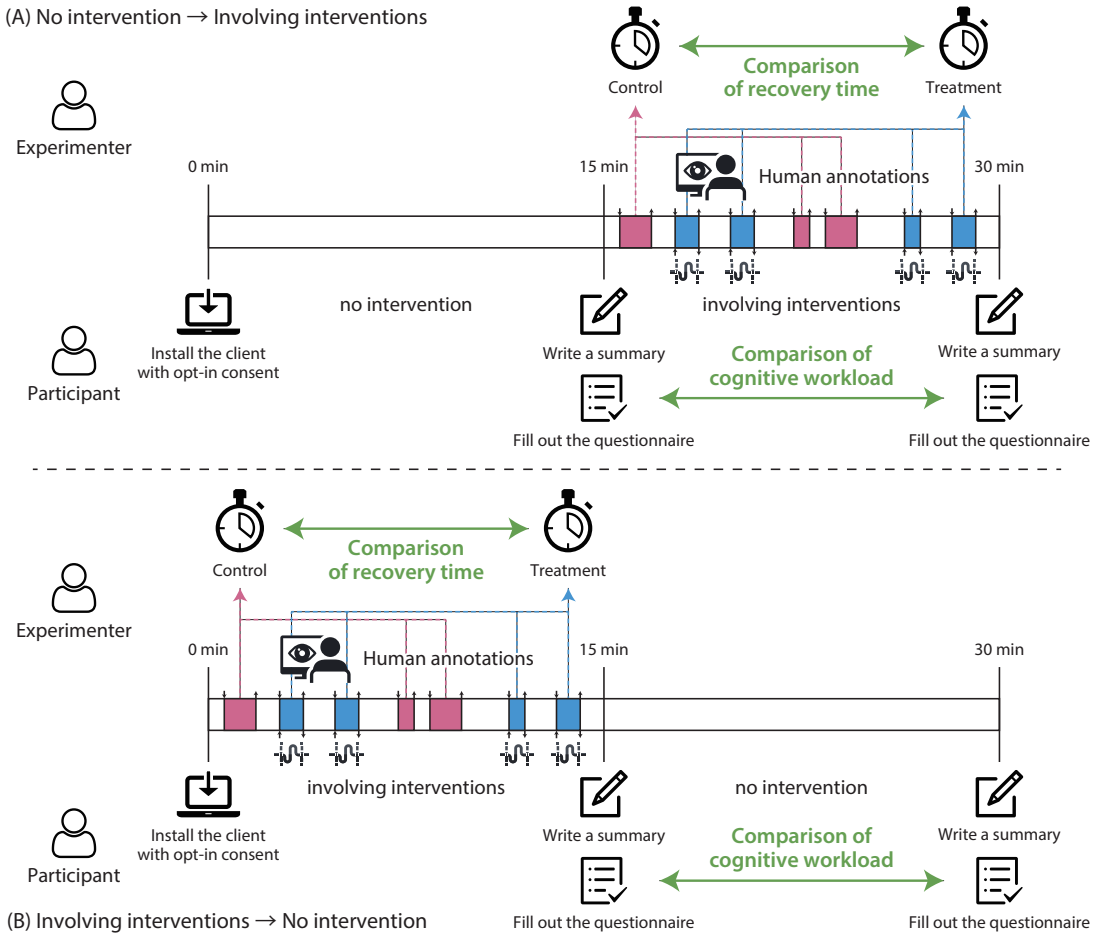


Figure 8: Example illustration of the procedure for the first experiment. (A) Half of the participants first experienced the part with no intervention and then experienced the part involving interventions, and (B) the others followed the reversed order.

a random basis. If the intervention did not consume the participant's conscious awareness, the scores of the two parts would not be significantly different. In addition, at the end of the session, the participants were asked for their comments about their feelings or anything they noticed. In total, the entire session took about an hour to complete.

4.4.6 Results

Recovery time As shown in Table 2, the mindless intervention significantly shortened the recovery time according to the unpaired t-test (Cohen's $d = 1.0044$, $p < 0.0001$). The distribution of the recovery time is shown in Figure 9, which also confirms this reduction. This result supports that Mindless Attractor helped participants refocus their attention.

I also investigated which of the four perturbation patterns (*i.e.*, halving or doubling the

Table 2: Comparison of the recovery time and cognitive workload score between the control and treatment conditions. The treatment condition involved the mindless intervention.

Measure	Treatment	Control	p-value
Recovery time	17.71 s (±10.52 s)	32.25 s (±16.92 s)	< 0.0001
Cognitive workload	26.00 (±10.32)	27.00 (±9.13)	0.5212

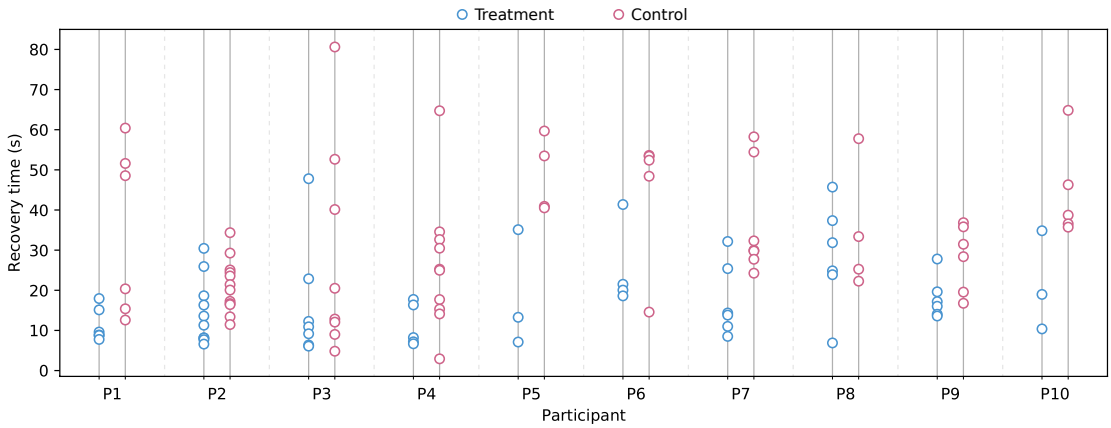


Figure 9: Distribution of the recovery time across each participant and the experimental conditions.

Table 3: Occurrence of the four perturbation patterns that were executed just before participants returned their attention. The comparison with the total occurrence suggests that there was no significant difference in effectiveness ($p = 0.2794$).

Perturbation	Halve the volume	Double the volume	Lower the pitch	Raise the pitch
Occurrence just before participants returned their attention	19	7	14	16
Total occurrence	50	47	50	55

volume and lowering or raising the pitch by one tone) effectively helped participants refocus their attention. Specifically, the last perturbation pattern before each time the participant returned their attention was examined to count their occurrence, as shown in Table 3. This examination is based on the assumption that the intervention just before the participant's attention returned is the cause of the change in the participant's state. According to the χ^2-test comparing with the total occurrence, the results were not significantly different, indicating that each pattern equally helped participants recover their attention (Cramer's $V = 0.1220$, $p = 0.2794$). In other words, it suggested no significant difference in the effectiveness of the four perturbation patterns.

Cognitive workload Also, no significant difference was found in participants' cognitive load scores according to the paired t-test (Cohen's $d = 0.2110$, $p = 0.5212$), as presented in Table 2. That is, it is suggested that Mindless Attractor did not consume participants' conscious awareness or at least did not negatively affect participants' cognitive load by consuming their conscious awareness. Thus, in combination with the effect on the recovery time, H1 was supported.

Comments Also, the comments that the participants wrote at the end of the experiment were examined. Here, three participants mentioned that they did not notice any intervention, although they were informed of the intervention beforehand. Interestingly, the recovery time for these three participants also showed a significant difference (Cohen's $d = 0.8105$, $p = 0.0122$) between the treatment (15.88 s on average) and control (28.34 s on average) conditions. Thus, it is suggested that the mindless intervention worked even when it was not noticed by participants, further supporting that Mindless Attractor did not consume the participants' conscious awareness. This point not only corroborates H1

but also shows consistency with the discussion by Adams *et al.* [1]. It was also notable that, although five participants mentioned that they noticed the changes in volume, no participant recognized the changes in pitch. That is, although no significant difference was found between the effectiveness of the four perturbation patterns in Table 3, their noticeability varied, suggesting room for investigation.

Nevertheless, no participants regarded the mindless intervention as disruptive or annoying. Rather, two participants made positive comments about it:

> I found it useful because it naturally brought my attention back to the video when I thought something might have changed in the speech. (P1)

> It was nice as it made me feel like...the computer was recommending me to concentrate, rather than warning me. (P4)

In particular, the latter comment suggested that the mindless intervention can mitigate the negative effect that might be caused by false-positive detection when combined with a sensing module based on an off-the-shelf ML model. These results led to conducting a second experiment to evaluate this possibility, as discussed in Section 4.3 when posing H2.

4.5 Experiment 2: Evaluation of H2

4.5.1 Design

To evaluate H2, another experiment was conducted in the same manner as Section 4.4, replicating a video-based learning situation. However, in this case, a sensing module using an off-the-shelf ML model was combined, instead of manually activating interventions, and the effects of the mindless intervention and the alerting approach were compared. Here, a within-participant design was used over three conditions: mindless, alerting, and control (no intervention). The control condition was added to confirm that the mindless intervention was at least effective in contributing to refocusing users' attention as an interactive system harnessing an ML-based sensing module. H2 is thus supported if the following two points are confirmed: Mindless Attractor helps participants refocus their attention, and participants favor Mindless Attractor over the alerting approach.

4.5.2 Measure

Similar to the first experiment, the time with regard to whether participants were paying attention was measured. However, this experiment introduced a different approach for evaluating the time factor (*i.e.*, total distracted time instead of the recovery time) as well as a measure for behavioral intention.

Total distracted time Although we have confirmed that Mindless Attractor can help participants return their attention, it is desirable to investigate whether the total time that they are distracted during video-based learning is reduced. In other words, it may be possible that, though the mindless intervention shortened the recovery time, the participants were distracted more frequently, especially when the mindless intervention was combined with an ML-based sensing module having a risk of false positives. To compute this metric, human annotations were collected for each participant, as in Section 4.4, and aggregated the duration when the participants were not paying attention. If the total distracted time in the mindless condition is significantly shorter than in the control condition, it is suggested

that Mindless Attractor can make users more likely to pay attention, even in combination with a sensing module based on an off-the-shelf model. It should be noted that, due to the false negatives of such a sensing module, there would be a case when the intervention is not triggered even when the participant is actually losing their attention and a case when the intervention is deactivated before the participant refocus. Therefore, calculating the recovery time as in Section 4.4.6 is not appropriate in this second experiment, further rationalizing the introduction of the total distracted time as a different metric.

Behavioral intention This metric was prepared to evaluate whether the mindless intervention was favored over the alerting approach. The concept of behavioral intention is guided by the Technology Acceptance Model [64], which explains users' attitudes towards technologies and is frequently used to evaluate how likely individuals are to use the technologies. Here, the questionnaire to measure behavioral intention was used in the same manner as the previous studies [310]. Suppose this score in the mindless condition is significantly better than that in the alerting condition. In that case, we can confirm that Mindless Attractor can be favored over the alerting approach, especially when integrated into an interactive system harnessing a sensing module based on an off-the-shelf model.

4.5.3 Material

Similar to the first experiment, a video recording of a 30-minute lecture on social sciences was prepared. Again, the experiment was conducted remotely, and the video was presented using Zoom's screen-sharing function. However, this second experiment required a system that automatically detects the status of participants' attention. To implement this sensing module, I followed previous studies that estimated participants' attentiveness based on their visual cues, as reviewed in Section 4.1.1. Specifically, the video stream of face images of each participant was analyzed by leveraging an ML model that can detect their head pose in real time. If the module detected that the participant was looking off the screen, the system judged that the participant was failing to pay attention to the video lecture and activated an intervention.

Figure 10 illustrates how the system processed the video streams of participants and intervened in them. Here, videos were processed on a frame-by-frame basis in the following manner. First, a human face was detected and located in each frame using an off-the-shelf model of RetinaFace [66]. Face alignment was then performed to obtain facial keypoints using another off-the-shelf model released by Fang *et al.* [79], which is known to estimate keypoints with high accuracy. Based on the estimated facial keypoints, the head pose was calculated by solving a perspective-n-point problem. Next, the estimated head pose was passed to the experimenter's PC, which checked whether the passed head direction was off-screen or not. The experimenter had conducted a calibration process beforehand to calculate the threshold for this judgment, in which participants were asked to track a red circle that appeared and moved along the edge of the screen. Participants were told to track the circle by moving their head, *i.e.*, not following it only by moving their gaze. The maximum head rotations for each direction (top–down and left–right) were calculated and regarded as the range where the head is toward the screen. In other words, when the estimated head pose was out of this range, then the system judged that the participant was looking off the screen, and thus, losing their attention.

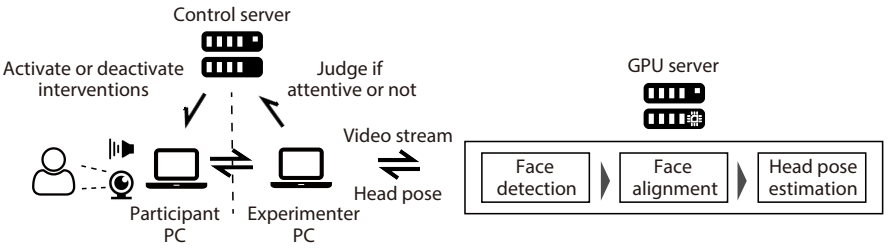


Figure 10: Architecture of the entire system implemented for the second experiment.

While the participants were watching the video, changes in their state—*i.e.*, whether they were looking at the screen or not—were shared with another control server maintaining a WebSocket connection with the client software. The control server then correspondingly sent activation or deactivation commands in the same manner as the first experiment. All of the above processes were performed in real time with a frame rate of 15 FPS.

In addition to the sensing module, an intervention to explicitly alert users was implemented in the client software to be compared with the mindless intervention. In this case, the client software played a short beep for 0.1 s, which followed the previous study's use of a beep alert [147], instead of perturbing the audio output. Once the alert was activated, it replayed the same beep every 3 s until it received the deactivation command, in the same manner as the mindless condition.

4.5.4 Participants

This experiment involved 20 participants, five of whom were female. They were recruited in the same manner as the first experiment. Eight of the participants participated in the first experiment, which had been held at least two weeks before this experiment. The participants were asked to prepare a PC in a quiet room and to enable their faces to be captured with a webcam, as in the first experiment.

4.5.5 Procedure

Similar to the first experiment, each participant experienced a session of watching the 30-minute video using a computer connected over Zoom. As before, the participants were told in advance that they would be asked to write a few sentences summarizing the video and also allowed to bring and use their smartphones in the session. As illustrated in Figure 11, each session consisted of three parts lasting 10 minutes each: one with no intervention, another with the mindless intervention, and a third with the alerting approach. The order of these three parts was automatically randomized among participants, as described later in this section. After each session, participants were asked to write a summary. They were also asked to fill out the questionnaire measuring behavioral intention when they finished a part with either the mindless intervention or the alerting approach. The scores between the two conditions were compared to examine which approach participants favored.

Before starting the first session, the experimenter performed a calibration process to determine the threshold for whether the participant's head pose was out of the screen, as described in Section 4.5.3. The experimenter explained that the participants should not move their PCs until the entire process was complete and advised them to find a comfortable position before the calibration process started. In each of the three parts, the experimenter

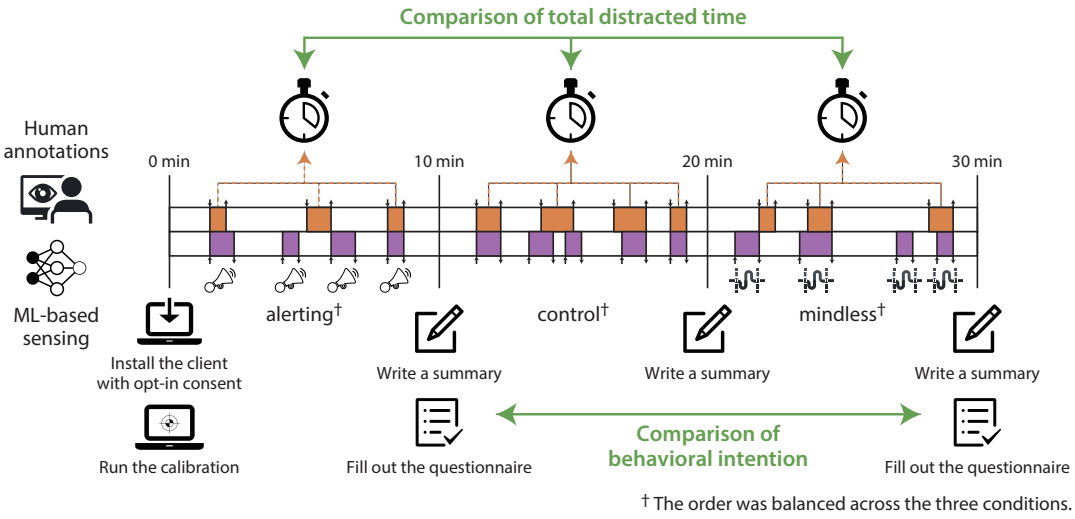


Figure 11: Example illustration of the procedure for the second experiment. Each participant was randomly assigned to one of six possible orders of the three conditions.

Table 4: Confusion matrix between the human annotations and the detection results of the ML-based sensing module in regard to participants' attentive state.

		Detection result	
		Attentive	Distracted
Human annotations	Attentive	435.4 min (68.5 %)	78.0 min (12.3 %)
	Distracted	51.4 min (8.1 %)	70.7 min (11.1 %)

manually annotated whether the participant was paying attention to the video lecture, similar to the first experiment. To avoid bias, the experimenter was blind to which of the three conditions had been applied to the participant. Specifically, the control server (see Figure 10) decided the order of conditions in each session, and the experimenter did not have access to this information until the session ended. The obtained annotations were used to calculate the total distracted time for each part, while the ML-based sensing module triggered interventions to the participants in either the alerting or mindless condition, as described in Section 4.5.3. In the alerting condition, participants were exposed to the beep sound when the system judged that they were losing attention, whereas they were exposed to perturbations in the speech in the mindless condition. In the control condition, the client system did not intervene. In each part, the sequence of the system's judgment was recorded along with timestamps, which was later used to assess the accuracy of the sensing module by comparing it with the human annotations. Finally, at the end of the session, the participants were asked for their comments about their feelings or anything they noticed. In total, the entire session took about an hour to complete.

4.5.6 Results

The experiment provided the following results.

Sensing accuracy I first examined the accuracy of the sensing module based on the off-the-shelf model in detecting participants' attentive state. Specifically, the human annotations and the detection results of the module were compared, yielding Table 4. As

a result, the accuracy across all the participants was 79.6 %, which was relatively close to the previous study [298] that achieved the accuracy of 82–85 % using only head pose. Note that the accuracy varied among participants (64.9–93.0 %), which implies that some environmental factors (*e.g.*, the distance to a camera or lighting conditions) might largely affect the detection results. At the same time, the sensing module exhibited a lot of false-positive detection, as its precision was 47.6 %, which suited the aim to investigate the effect of Mindless Attracter while having a risk of false positives.

Total distracted time Next, based on the human annotations, the total distracted time was calculated for each participant, as presented in Figure 12. A significant difference was found among the three conditions according to ANOVA ($F(2, 57) = 8.5773$, $\eta^2 = 0.2313$, $p = 0.0005$), and thus conducted a post-hoc test. As a result, the control condition showed significant differences against the mindless and alerting conditions (Cohen's $d = 1.1795$, $p = 0.0013$ and Cohen's $d = 1.0828$, $p = 0.0032$, respectively). On the other hand, no significant difference was found between the mindless and alerting conditions. From this result, it was confirmed that Mindless Attractor is an effective means to refocus users' attention even when combined with an ML-based sensing module, as the mindless condition significantly reduced the total distracted time than the control condition. In addition, it is notable that Mindless Attractor would work effectively as well as the conventional alerting approaches since the mindless and alerting conditions showed similar distracted times.

I also examined how many times the participants got distracted because it was possible that the mindless interventions increased the frequency even though the total distracted time was reduced. As shown in Figure 13, no significant difference was found among the three conditions ($F(2, 57) = 0.1796$, $\eta^2 = 0.0062$, $p = 0.8360$). It can be explained as follows: the participants were almost equally likely to lose focus in all three conditions, but if there was an intervention, they often refocused their attention to the video earlier, as confirmed in the first experiment. As a result, their distraction time in the mindless and alerting conditions was significantly reduced than the control conditions. From these results, we can conclude that H2 was supported in terms of the effectiveness of Mindless Attractor.

Behavioral intention Lastly, participants' scores of the behavioral intention were compared between the mindless and alerting conditions. As presented in Figure 14, a significant difference was found (Cohen's $d = 0.7025$, $p = 0.0054$) according to the paired t-test. This means that, compared to the alerting approach, the participants showed their stronger intentions to use the implemented system when it is combined with the mindless intervention. This result supports that Mindless Attractor is much preferred by users than the alerting approach, as hypothesized as H2.

Comments The above results coincided with H2, confirming that Mindless Attractor helps participants refocus their attention and is favored over a conventional alerting approach. In addition, the comments obtained at the end of the experiment corroborated H2, especially in regard to the unacceptability of the alerting approach.

> I felt like the beep sound made me lose focus. It was frustrating, especially when I was concentrating. (P9)

> The beep felt like noise because it overlapped the speech though I wanted to listen to what was being said. As a result, my concentration was more disrupted

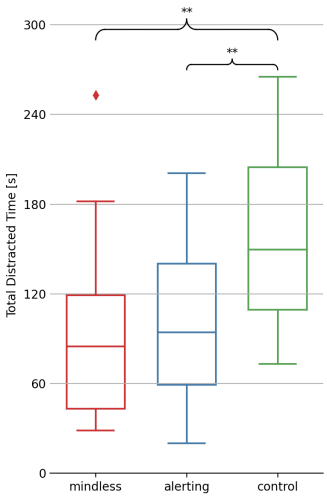


Figure 12: Comparison of participants' total distracted time. A significant difference was found between the control condition and the other conditions.

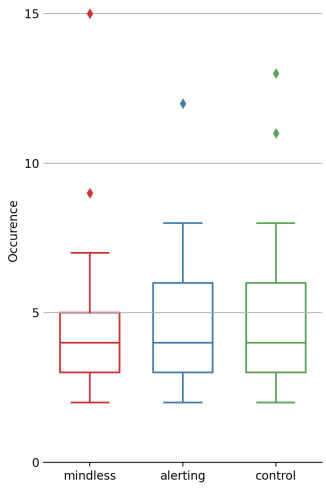


Figure 13: Comparison of how many times participants got distracted. No significant difference was found between the three conditions.

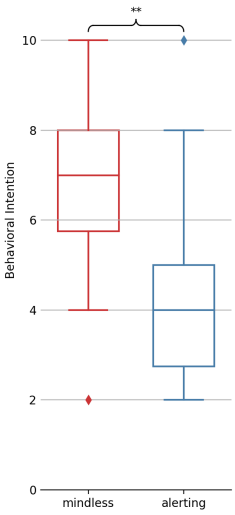


Figure 14: Comparison of participants' scores of the behavioral intention. A significant difference was found between the mindless and alerting conditions.

> than the case that I had not used the system. (P12)
>
> I thought the one with the beep sound might be a good signal until halfway through, but then it came to ring repeatedly even though I was concentrating. As a result, I stopped caring about the sound. (P2)

These comments confirmed the anticipation that explicitly alerting users based on false-positive detection makes them distracted and frustrated, which can lead them to ignore the intervention. In addition, one participant suggested that such negative effects can be caused even when the intervention was activated by accurate detection:

> I was disgusted by the alarm, which rang when I was using my smartphone for googling a word I never heard. (P8)

In contrast, the mindless condition was totally favored, as follows:

> In the part [of the mindless condition], I felt like I was able to focus on the lecture relatively well. (P12)
>
> I did not notice much of a change in the audio, but when I compare the three parts, I seemed to be able to maintain my concentration the most. I think having such a system that brings back my attention without making a big deal will help me stay focused in usual situations. (P3)
>
> When the pitch of the speech became higher, I paid attention to the video as I felt strange a little. It did not provide a sense of being angry, compared to the beep alarm. (P11)

These comments corresponded to the comparison of the scores of behavioral intention (Figure 14).

Furthermore, 17 of 20 participants agreed they often have trouble maintaining their attention and computationally solving it would be beneficial, like:

> I find it difficult to maintain my attention in such online situations because of the lack of eyes around. (P1)

In addition, they suggested that the presented interaction design can be used outside video-based learning situations.

> I thought it would be nice to be able to introduce a similar system in offline situations. I will appreciate it if some device such as a smartwatch helps me refocus when I am losing my attention from an important conversation. (P4)

The obtained comments not only corroborated the effectiveness of Mindless Attractor through supporting H2 but also highlighted the further potential of the presented interaction design.

4.6 Discussion

So far, by verifying H1 and H2, I have demonstrated that Mindless Attractor works effectively as a novel interaction design to support users' participation during video-based learning. In this section, I contemplate the findings of the experiments, envision future application scenarios, and discuss limitations and future directions in terms of paving the way for supporting users in video-based communication.

4.6.1 Necessity of mindless intervention in machine-learning-based interaction design

The results of the second experiment supported H2, as participants favored the mindless intervention, while the alerting approach was not accepted. Specifically, the obtained comments suggested that participants were annoyed by the alerts when they were triggered by false positives of the sensing module based on the off-the-shelf ML model. In other words, mistakenly intervening in an explicit manner while users are concentrated on the primary task can unnecessarily consume their conscious awareness and eventually disrupt their experience. Indeed, such failures in designing interactive systems based on ML-based sensing modules have been pointed out in a guideline for human-AI interaction [6]. That guideline emphasized the importance of considering that such AI-infused systems may demonstrate unpredictable behaviors due to false positives and false negatives. Consequently, it suggested that an effective approach in designing AI-infused systems is to enable users to dismiss the undesired functions instantly.

From this guideline, we can learn two points. First, the participants' acceptance of the presented interaction design can be attributed to the fact that it indirectly follows this guideline by not consuming users' conscious awareness, letting them not mind the mistakenly triggered interventions without much cognitive workload. This suggests that the design of Mindless Attractor can be applied to various cases as an intervention method when integrated with ML models. Particularly, when we consider the use of off-the-shelf models, this design would be a clue to finding a subproblem in user needs that is tractable by the models. Second, such an existing guideline for human-AI interaction can also be applied when we construct interaction design using off-the-shelf models. This implies that

the strategies presented in Section 3.2 operate in parallel with existing guidelines and complement them.

4.6.2 Limitations

Though the experiments have demonstrated that Mindless Attractor is a promising interaction design, there are some limitations. Initially, further investigations involving a greater number of participants and diverse lecture content are desirable to generalize the results. For example, if a lecture is so attractive that learners are not distracted from the video, the mindless intervention would not be necessary even though, at worst, it would not be harmful, as its impact on cognitive load was not observed in Section 4.4.6. Secondly, the experimental design was based on the discussion of Mindless Computing [1], considering users whose motivation for obeying the intervention is not always assumed. Specifically, the design was intended to nudge the participants not to be much motivated to watch the video, like allowing the use of smartphones, and skipped the measurement of the participants' motivation in the experiments. However, this means that their results would not necessarily guarantee the universal effectiveness of the intervention for users with any level of motivation. Thus, evaluating participants' motivation and exploring its correlation with the efficacy of Mindless Attractor can be intriguing.

Furthermore, refining the design of alerts can mitigate the negative impact suggested in the second experiment. While a simple beep was used as an alert, alternative methods to inform users in less annoying manners are possible. In particular, Weiser and Brown [326, 327] conceptualized "calm technology" as a more acceptable communication channel from computers. For example, alerting users with less explicit sounds (*e.g.*, birds chirping) could be preferred to a simple beep sound. In addition, if we ignore the requirement of using the auditory modality, showing a status lamp on display is an alternative to informing users that they are losing attention. However, as Adams *et al.* [1] pointed out, these techniques require users' conscious awareness (*e.g.*, interpreting the status based on the lamp) to induce behavioral change, while mindless computing does not. Therefore, Mindless Attractor can be differentiated from alerting approaches in that it can work without consuming users' conscious awareness, as suggested in the first experiment (see Section 4.4.6). That said, it is desirable to explore sophisticated alerting approaches to draw further implications in comparison to the mindless intervention.

At the same time, the design of the mindless intervention has also room for exploration. Currently, as explained in Section 4.2.2, the perturbation of the pitch or volume of the voice is employed based on the nature of human speech communication. Though the results were not statistically examined in a participant-wise manner due to the small number of perturbations activated for each participant, there were individual differences in terms of their effectiveness, which would imply the possibility of personalizing the intervention patterns. Moreover, human brains are known to show a special response to a self-voice [57] or a familiar voice [21]. Thus, a possible intervention might involve computationally modifying a voice so as to be similar to a self-voice or familiar voice when users are not paying attention. This can be achieved through existing techniques for high-fidelity real-time voice conversion [299, 8].

It would also be interesting to investigate whether the presented interaction design

that helps learners pay attention contributes to their learning performance. Considering that previous studies adopting explicit feedback to help learners pay attention have shown a positive impact on performance [15, 337], the mindless intervention can be expected to have a positive effect. This is because the mindless intervention exhibited an effect in reducing distracted time while showing no significant impact on the cognitive load in Section 4.4.6. Examining the long-term effect of the presented interaction design is also intriguing. Though this design is based on the concept of Mindless Computing, which Adams *et al.* [1] have described as having long-term effectiveness, it is difficult to deny, without further investigation, the possibility that users will become acclimated to the perturbations. However, even in this case, the combination with voice conversion could be a remedy, as it enables as many patterns of interventions as the number of conversion targets.

4.7 Summary

This chapter presented a novel interaction design, Mindless Attractor, which helps users refocus their attention in a mindless manner. It was guided through considering how to apply off-the-shelf ML models to such a complex user need and exploiting human cognitive processes, as stated in Section 3.2. Specifically, the interaction design leverages the nature of human speech communication and perturbs the voice that users hear when they are detected as losing their attention by an off-the-shelf model. The effectiveness of Mindless Attractor in a video-based learning context was confirmed in two-stage experiments: the first experiment showed that it helped users refocus their attention without consuming their conscious awareness, and the second experiment supported the effectiveness when implemented as an interactive system with a sensing module using an off-the-shelf model. The results partially answer the RQ1 and RQ2 of Section 3.4. We confirmed that Strategy #1 can be used to construct a new interaction design for video-based learning situations while using an off-the-shelf model, and the interaction design effectively addressed the user need to maintain their attention.

5 CatAlyst: An interaction design for preventing procrastination in intellectual tasks with off-the-shelf large generative models

Given that many office workers perform intellectual tasks, such as writing, coding, and editing slides, using computers every day, this chapter focuses on supporting intellectual tasks using interaction design harnessing off-the-shelf ML models. During an intellectual task, workers are known to often experience a lack of progress and stay away from it because of the high cognitive load required to perform the task [188, 225, 314]. Such behavior often leads to procrastination, causing stress and degraded self-efficacy [163, 83, 257, 22]. In this context, several approaches to increasing workers' engagement via intervention have been proposed in the HCI literature. For example, presenting an encouraging message or feedback is one useful approach in persuasive technology [25, 311, 180]. Previous studies have suggested that context awareness is important for effectiveness [247] and variation in the intervention content is no less essential for its continued effectiveness [156]. In this chapter, I would like to show that we can provide more context-aware, variational content by using ML, while these approaches did not employ ML technologies.

Still, preparing a tailored ML model for this purpose is troublesome due to the reasons presented in Section 2.3. Therefore, based on Strategy #1 in Section 3.2, I explored the way to apply off-the-shelf models by taking human cognitive processes into consideration. Here, the elaboration likelihood model (ELM) [230] was specifically leveraged. ELM posits that, for a persuasive message to lead to behavioral change, two key elements must be present: *motivation* and *ability*. This means that, if a user is highly motivated and has the ability to process the message, it can lead to behavioral change. Conversely, if the motivation is low or the ability to process is insufficient, behavioral change is unlikely to occur. Fogg [82] elevated this concept within the context of persuasive computing, proposing a model where an appropriate *prompt* from computers at the right moment is also a crucial requirement. It implies that, even with an off-the-shelf ML model that is not tailored specifically for behavioral changes regarding task engagement, interactive systems can contribute to engaging in intellectual tasks if we can present prompts that enhance motivation and ability using the model.

Based on this idea, I present a novel interaction design, *CatAlyst*, that harnesses a large generative model that is shared as an off-the-shelf ML model. This aims to lower the cognitive load necessary to resume working on intellectual tasks, as *catalysts* in chemistry, via prompting workers under the assumption that the lowered cognitive load leads to eliciting their motivation and instilling a sense of capability in resuming the tasks. Specifically, CatAlyst first detects when workers' progress is halted and then intervenes with them by presenting the continuation of their work, which is generated by off-the-shelf generative models. Here, it might be tempting to think that using large generative models to automate intellectual tasks could directly address the root cause of user needs. However, the efficiency and subsequent workers' satisfaction with such automation approaches depend on the accuracy and fidelity of the ML models. For instance, trust may not be built between ML-based systems and workers because of poor outputs that conflict with workers' intentions or do not meet their expectations [70, 144, 359, 9], resulting in ignorance of the systems.

For example, Yang *et al.* [352] proposed an ML-based system that supports the writing of fictional stories through text generation. However, it was suggested that the output of the model might irritate users due to its low accuracy, lack of usability, and insufficient transparency. In addition, assuring accuracy for each of a variety of task domains by customizing models is challenging, as explained in Section 2.3. On the other hand, CatAlyst aims to retrieve workers' interest and encourage them to resume the original task by lowering the hurdle, not contributing directly to workers' tasks with generated content. Since workers recognize CatAlyst as an intervention to encourage their work rather than proactively using ML technologies with high expectations, they would not feel as irritated as systems like the above one [352] even when generative models perform poorly.

To test the effectiveness of this interaction design, a series of evaluations was conducted. First, a prototype running on a Web page was developed to support writing tasks by generating a continuation of sentences in progress using GPT-3 [35], and its effectiveness was confirmed by conducting a user study with 12 participants. Second, to examine how CatAlyst can be used in an unconstrained situation for a longer period, another user study was conducted, in which ten participants used the prototype for five days. Here, the purpose of the participants to use the prototype was varied, such as keeping a diary, writing a novel as their hobby, and writing media articles as their job. Finally, to demonstrate the extensibility of CatAlyst in different tasks, a second prototype was developed in the form of a Chrome extension that supports slide-editing tasks by generating the continuation of slides in progress using GPT-3 [35] and a diffusion model [250]. These studies showed that, under the condition that off-the-shelf generative models have a certain degree of generality but not optimal accuracy, it is possible to support workers in multiple tasks by lowering the hurdle for resumption. In other words, this is another example that the interaction design informed by the strategies in Section 3.2 effectively addressed the user need by leveraging an off-the-shelf ML model.

5.1 Related work

To situate the presented interaction design, I first review the research trend of leveraging generative models to improve worker productivity while performing tasks, which leads to the discussion about why supporting workers' task engagement is specifically important. Also, factors affecting workers' task engagement were discussed based on existing HCI studies to illustrate the novelty and challenges of utilizing generative models for this purpose.

5.1.1 Generative models for improving task efficiency

As many tasks have been digitized, there is a strong need for computational support to improve the efficiency of such tasks, producing many software systems designed for various tasks, such as Microsoft Office 365, Adobe Creative Cloud, *etc.* Such systems have now started to incorporate ML technologies to add machine intelligence to accelerate workers' task performance. For example, Gmail [49] can automatically generate response candidates as well as complete sentences. GitHub Co-Pilot [375] can generate programming code based on the natural language description of the functionality provided by users. These products demonstrate the enormous potential of leveraging ML technologies to directly contribute to and replace parts of human tasks, significantly improving workers' task efficiency.

More features and interactions have been proposed at the research level using ML-based generation models. One of the most prominent examples is writing support systems that leverage large language models [35]. There are a number of concurrent studies that demonstrate how human writers and ML-based generation models can collaborate to support writers [166, 260, 63], such as in creative writing [248, 265, 352, 362, 275]. For example, Dang *et al.* [63] proposed a writing support tool that provides paragraph-wise summaries as self-annotations while workers write the content. They found that the generated summaries gave the workers an external perspective on their writing and helped to polish the content. A series of ML models have also been developed for tasks other than writing, such as slide-editing [263, 169, 371], music creation [182, 287], and drawing [173, 150, 368, 367]. For example, Sefid *et al.* [263] proposed an ML-based slide generation method from PDF files for scientific papers, whereas Zheng *et al.* [371] proposed an approach of converting programming notebooks into slides and showed that it improved the efficiency of a slide-editing task.

While these studies show promising results of employing ML-based generation models to improve task efficiency, there are two gaps: the wall of accuracy and the dependency on workers' engagement levels. First, insufficient accuracy and fidelity in the output of ML-based systems can devastate trust in the systems, leading to ignorance of them. Yang *et al.* [352] reported that poor quality in generated texts could irritate fiction writers. Kim *et al.* [150] found that poor quality of output from AI does not provide rich inspiration for workers. It would be ideal to customize ML models for each worker so that they can achieve sufficient accuracy in individual tasks [278], but this is not necessarily feasible, especially because of the huge cost of training and maintaining large models, as discussed in Section 2.3. Furthermore, Yang *et al.* [356] pointed out the existence of inherent uncertainty surrounding their capabilities and output complexity, making it difficult to construct trustful systems, even with such customized models. Given these inherent concerns, it is no less important to study generalizable approaches as to how off-the-shelf ML models can contribute to workers performing various tasks without individual model tuning.

Second, the above studies implicitly assumed situations where workers are amenable to using the systems and continue working on tasks, which might not always hold in actual cases. In fact, as described in the next section, workers frequently struggle to maintain focus while performing tasks for various reasons, including the required high cognitive load and multitasking. According to ELM [230], when not only their attention is away from the task but also their motivation is low, workers might not fully benefit from the systems even if they feature more useful functions. This point requires an alternative approach that targets workers who may not be fully immersed in their task, in contrast to previous studies that aim to improve the productivity of workers who already engage in the task.

5.1.2 Needs for supporting workers' task engagement

Here, the need to support workers' task engagement is emphasized by the fact that modern knowledge workers need to concentrate on a task for a long time, such as writing, coding, and slide-editing. Particularly, it is difficult to maintain concentration for a long time because these tasks cause mental fatigue, as human concentration is known to break after 30–45 minutes when performing these tasks [188, 225]. It is natural for people to become

distracted, and this tendency is even higher in remote work scenarios [314], which have increased significantly in these years owing to COVID-19 [228]. Furthermore, since such tasks require a high cognitive load, workers often experience difficulty, particularly when starting tasks or resuming after interruption. For example, Mark *et al.* [192] revealed that it could take 25 minutes to return to a task on average once distracted from the task.

This further leads to procrastination [163, 83], which is known to be a major problem for office workers today, as it can lead to stress and problems in relationships, work, and health, such as depression [257, 22, 52]. However, it is known that the causes of the high hurdle to starting work are complex [319]. One of the major causes is perfectionism among workers, leading to their assumption that tasks must be completed flawlessly. They often stop thinking before starting tasks because they consider too much about the expected workload and sometimes feel fear of failure [261]. In response, practitioners often emphasize the importance of starting tasks anyway, for instance, for 3–5 minutes before deeply thinking about the workload [239], as one of the best ways to engage in the tasks. The Pomodoro technique [55] is a strategy to divide the entire task into small segments so that workers can start working on one of them with a lowered hurdle. The presented interaction design is informed by such practices based on human cognitive processes, as it uses off-the-shelf ML models to guide workers to take the first step to start working when they are not immersed in the task, as described in Section 5.2.2.

5.1.3 Interventions for supporting worker's task engagement

Meanwhile, there are several approaches proposed to support workers' task engagement through interventions when they are distracted. Site blockers are a typical example of explicit intervention since they force users to return to their intended task by blocking access to specific entertainment Websites. However, this is a primitive and static approach that increases the psychological burden on the user owing to its coercive power [156], resulting in workers uninstalling the systems. Nudge has also been extensively studied [189, 180, 311, 247, 26, 25] to induce behavioral changes. For example, encouraging messages have been popularly employed to motivate students [25] or induce exercise behavior [311]. Furthermore, Liu *et al.* [180] proposed a method to motivate workers to return to the task of editing ill-formatted documents by designing feedback reflecting their progress based on the perspective of persuasive technology. Their user study showed a significantly shorter average off-task time by adding visually encouraging feedback that reflected their progress. Similarly, Rodriguez *et al.* [247] suggested presenting context-aware content as a factor for successful nudge design, as it improves personal relevance and motivation. Even though such nudging approaches can also be effective in our case, the inherent nature of nudging would not always be suitable in the context of procrastination on intellectually demanding tasks. This is because nudge has emphasized the importance of providing all available choices to people, without forbidding any of them [297], from the perspective of libertarian paternalism. Meanwhile, people who procrastinate intentionally choose to do so while recognizing the necessity of completing the tasks [163, 83]. Given this, the effectiveness of such nudge-based approaches is uncertain for our situation, where the target tasks may require higher effort and focus than those in prior studies, such as editing ill-formatted documents.

Here, the presented interaction design of providing the continuation of interrupted work to lower the hurdle for resuming a task can regarded as a sophisticated version of these nudge-based approaches. This leverages off-the-shelf generative models for this purpose since concurrent large generative models are adept at producing context-aware content, although they are sometimes unable to satisfy workers who expect the generated content to replace parts of their tasks directly [352, 150]. This point allows us to anticipate that such generated content could be used to support workers' task engagement by drawing their diverted attention and offering a base to work on, considering human cognitive processes such as ELM [230] and Fogg's model [82]. More specifically, we can expect that workers can resume the interrupted task with less effort if they find the generated content supporting their ideation or providing relevant information. I believe that this is another illustrative example showing how the consideration of cognitive processes enables a new application of off-the-shelf ML models.

5.2 CatAlyst

This section describes how CatAlyst intervenes with workers using the generated continuation of interrupted work. Also, the strategy behind this idea is elaborated from the perspective of human cognitive processes and persuasive technology.

5.2.1 Architecture

Figure 15 shows an overview of the presented interaction design for intervening with workers to enhance task engagement. CatAlyst first detects the moments where the progress of the workers' tasks is stopped based on the interaction log (A). It then generates content that is likely to follow interrupted work using an off-the-shelf generative model (B). For example, GPT-3 [35] can be used in writing tasks to predict what would come next. Finally, it sends notifications to the workers while showing part of the generated content to induce them to resume tasks (C). To expand the applicability, the pipeline is designed to be capable of using any large generative models as long as they can produce a context-aware continuation of the interrupted task. As mentioned at the beginning of this chapter, two prototypes for writing and slide-editing support were prepared based on this design. Both prototypes run as Web applications.

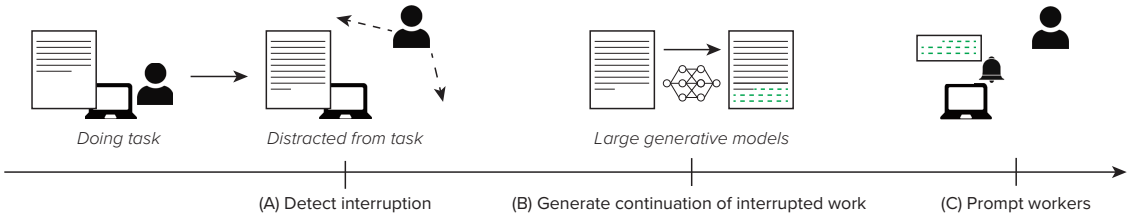


Figure 15: Overview of CatAlyst. (A) Detecting interruption based on workers' progress. (B) Generating continuation of interrupted work using off-the-shelf generative models. (C) Prompting workers to resume the task via notification.

For detecting a break in workers' concentration, previous research has explored several approaches, including image-based sensing using a webcam [11], and estimation from interaction log [180]. The prepared implementations adopted the estimation method using interaction logs to avoid the additional cost and risk of using sensors (*e.g.*, privacy).

Specifically, the prototypes send the interaction logs to a server and find a moment when the workers have not made any interaction for T s. The initial value of T was determined in a pilot study as the approximate time that distinguishes the actual distraction of the workers from the random interval during an engagement, which was 45 s. I was aware that, even if workers do not make any interaction over T s, they might engage in tasks, such as long contemplation, resulting in false-positive detection. Simultaneously, it was presumed that intervening with them in such moments using the generated continuation would not work negatively, as it could provide them with new perspectives or information that could facilitate their contemplation. This idea of designing an intervention that might work (or at least does not cause negative experiences) even when triggered by false-positive detection was common to Mindless Attractor, which was presented in Chapter 4. Note that we can make the value T controllable to provide them with global control of ML-based systems for a better human-AI relationship [6].

For generating the continuation of the interrupted work, CatAlyst uses off-the-shelf generative models, such as GPT-3 [35]. In the writing task, for instance, the text in progress was provided for a model to predict the text to follow. The same scheme can be applied to the slide-editing task, in which all text fragments and image captions are extracted from the slides in progress and provided for a model (Figure 16). Subsequently, the continuation prediction from the model is used to fill the new slide that is automatically appended. Here, CatAlyst can be applied to other domains by alternating the off-the-shelf model to be used with, for example, those for music creation [182, 287] and drawing [173, 150, 368, 367]. Furthermore, the slide-editing task illustrates how this interaction design of harnessing off-the-shelf generative models can be extensively applied to new tasks by formulating appropriate inputs (*i.e.*, prompt programming [244]) even if those tasks have not been considered at the time of training.

The prototypes notify workers using a Notification API of the browsers once the continuation is generated. It aimed to draw not only workers' attention but also their interest, as discussed later in Section 5.2.2. Therefore, the notification was designed to present the beginning of the generated continuations (*e.g.*, the first sentence of the generated text or the title of the generated slide). Note that CatAlyst can be extended to send notifications in other forms, such as using a smartphone to stimulate workers whose interest is deprived by smartphone apps.

5.2.2 Design consideration

Resuming an intellectual task while a worker gets distracted is expected to require more cognitive load than in cases where a nudge intervention is successful. In these situations, it is expected that encouraging feedback, as in conventional methods [180, 311], has room for improvement to provide more persuasive interventions and that prompting workers along with a continuation would be more effective. By doing so, CatAlyst encourages workers to start thinking about the task regardless of their status, even for a short time, which is known as a popular practice to engage in the task [239]. This idea was originally inspired by a practitioner's insight, which has been popularly adopted by readers. James Clear [56] proposed a strategy for behavioral change by employing an analogy of *activation energy* in chemistry. He explained that every habit has activation energy that is required to get

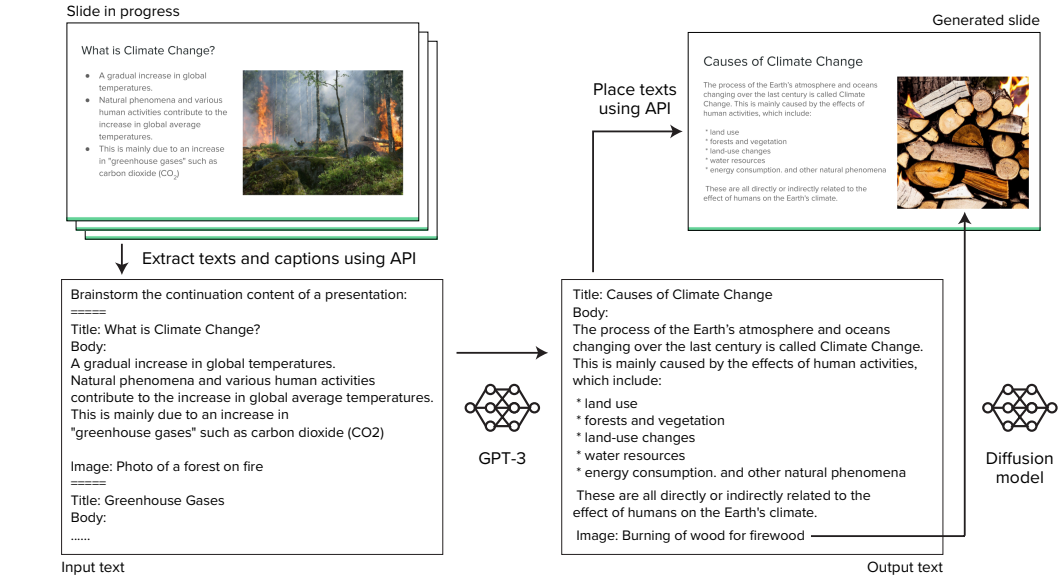


Figure 16: Actual example of how CatAlyst can generate the continuation in the slide-editing task using GPT-3.

started, and thus, it is important to introduce a *catalyst* to lower the activation energy. He further suggested utilizing intermediate steps with lower activation energy to induce behavioral changes step by step. Similarly, the presented interaction design of achieving such a *catalyst*-like effect using the generated continuation of interrupted work as the intervention can be broken down in detail, as shown in Figure 17.

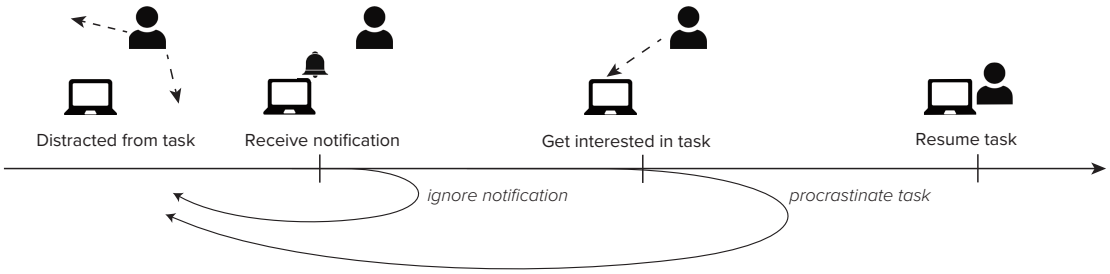


Figure 17: Interventions lead to workers' task resumption through multiple steps. At each step, there is a chance for an intervention to fail to induce workers' behavioral change.

Firstly, it is important that workers become interested in interventions and do not ignore them. Regarding this point, interventions via notification can at least draw workers' attention but do not always draw their interest. This means that it is easy to anticipate that workers will ignore them if the content of the intervention is not very useful [285] or attractive, as we often ignore push notifications on smartphones (the first failure case in Figure 17). Here, Kovacs *et al.* [156] showed that adding variety to intervention is effective in maintaining novelty. According to Berlyne [24], inducing curiosity is important to drive motivation. Wickens [330] emphasizes the importance of inherent interest based on a computational model on top of cognitive science. Given that, CatAlyst attempts to make workers interested in the content of the intervention by presenting the generated continuation of their interrupted work, which is more context-aware information, leading to the inherent interests of workers.

The second part of the intervention flow (Figure 17) is to induce workers' behavior of

resuming the task after they retrieve their interest in the original task. More specifically, CatAlyst aims to prompt them to resume the task quickly for 3–5 minutes to avoid procrastination and engage in it. As mentioned at the beginning of this chapter, according to ELM [230] and Fogg's revised model [82], behavioral changes are achieved when the elements of motivation, ability, and prompts are aligned. Regarding motivation, if workers perceive CatAlyst as a collaborator, they will gain motivation to utilize the generated content, as demonstrated in prior studies on human-AI collaboration [368, 367]. Regarding ability, the generated continuation of the interrupted work can be a help, for example, by supporting their ideation or providing reference information, as in prior studies about human-AI co-creation reviewed in Section 5.1.1. Therefore, it can be expected that this interaction design guides workers to resume the original task with enhanced motivation and ability, preventing procrastination.

5.2.3 Hypothesis

The design consideration of CatAlyst, as discussed thus far, poses the following hypotheses. First, it should be examined whether workers can recover their interest by being presented with the generated content as an intervention, as this is the first step toward resumption (see Figure 17).

> **H1**: CatAlyst is an effective means to keep attracting the interest of workers who are away from the task by presenting the continuation of interrupted work as an intervention.

Next, even if the workers' interest returns to the task by seeing the generated continuation, they might not resume the task and go back to the distracted states (*i.e.*, procrastination), given the required hurdle for working on the task. Therefore, it requires further investigation into whether the workers prompted by the intervention change their behavior (*i.e.*, resume the task), guiding us to posit the following hypothesis.

> **H2**: CatAlyst can induce workers' behavior to resume the original task effectively through the intervention.

If these hypotheses are supported, the effectiveness of CatAlyst in helping workers' task engagement will be validated. It is further anticipated that the overall productivity of workers would be enhanced with the intervention, leading to the following hypothesis. If this hypothesis is supported, CatAlyst can be used as an alternative approach to the previous studies to improve task efficiency.

> **H3**: CatAlyst can improve worker productivity by helping them avoid procrastination while performing tasks.

Finally, we can expect that CatAlyst can contribute to improving the subjective experience of workers who often face a high cognitive load during tasks. If this hypothesis holds true along with the above hypotheses, we can conclude that the design of CatAlyst functioned as intended. Therefore, this point leads to examining whether CatAlyst actually lowers the hurdle for workers to resume tasks and offers a favorable experience.

> **H4**: CatAlyst can lower the cognitive load imposed on workers while performing a task, thereby being favorably accepted by them.

5.3 Study 1: Writing Task

To test these hypotheses, I first conducted a user study in a controlled setting in which the prototype of CatAlyst for supporting workers' writing was used. This is because, while writing is one of the tasks that knowledge workers regularly perform, it requires a high level of skills and cognitive load, and workers often lose focus while performing it [198].

5.3.1 Design

A within-participant study was conducted to compare three conditions: *proposed*, *control*, and *none*. Participants experienced interventions from the prototype of CatAlyst in the *proposed* condition. In the *control* condition, they instead experienced a conventional intervention for supporting task engagement, that is, encouraging messages (see Section 5.1.3). In the *none* condition, they did not experience any intervention and only performed the writing task as usual.

5.3.2 Task

In this study, participants were asked to write essays of approximately 1,200 characters in Japanese. Six essay tasks were prepared based on publicized university exams to balance their difficulty, such as: "There are more people who eat alone these days. What do you think about this trend from the perspective of modern social structure?" Three were randomly chosen for each participant and assigned to the three conditions. It took roughly 30 min to write one essay.

5.3.3 Implementation

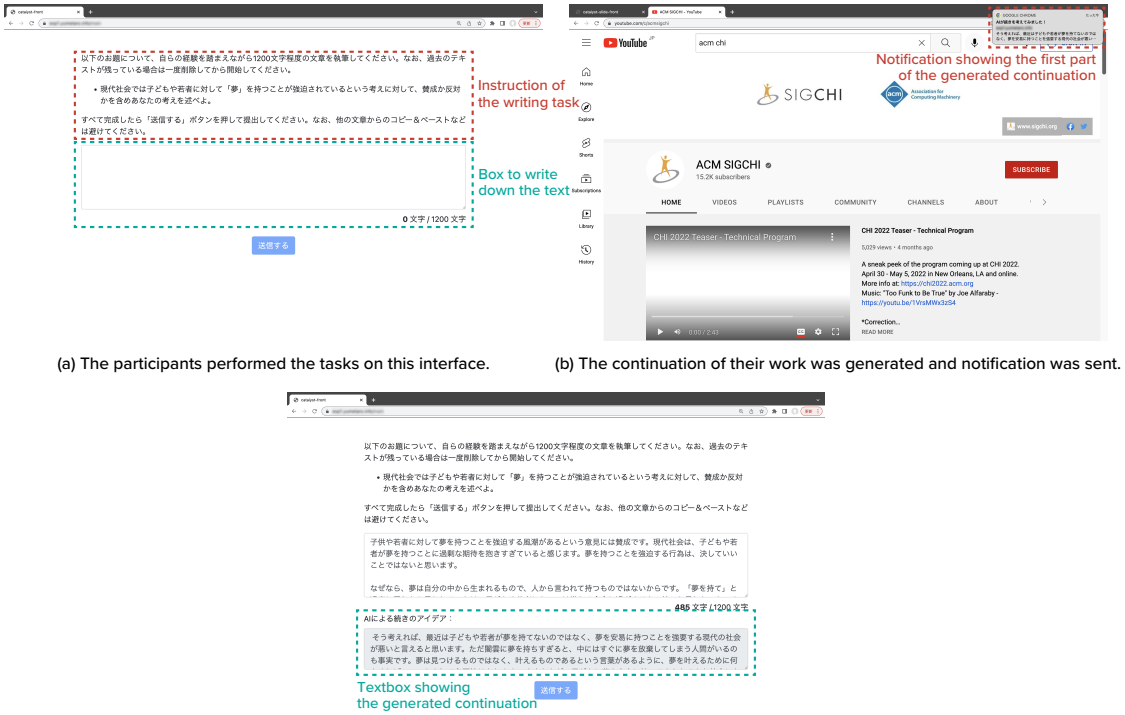
For this user study, three types of Web pages were prepared corresponding to the three conditions. As shown in Figure 18, they had a center box in which the participants wrote texts. In the *proposed* and *control* conditions, the Web pages recorded the interaction log of the participants and detected the moment they got distracted. In the *proposed* condition, the written text in progress was sent to a remote server, where its continuation was generated using the off-the-shelf GPT-3 model [35] that was pretrained with the C4 [240] and CC100 [58] datasets.[4] The Web page prompted the participants with a notification displaying its first sentence once the continuation was generated. In the *control* condition, the Web page did not send the text and immediately showed a notification displaying an encouraging message that was randomly chosen from the six patterns prepared based on a public site blocker.[5]

5.3.4 Procedure

The participant recruitment was conducted by publishing a Web page and, to secure the number of participants, posting a link to the page on social media. As a result, twelve participants (six males and six females, 24–59 years old) participated in the study. Each two of them were assigned to one of the six possible orders of the three conditions. The

[4] https://huggingface.co/rinna/japanese-gpt-1b

[5] https://chrome.google.com/webstore/detail/blocksite-block-websites/eiimnmioipafcokbfikbljfdeojpcgbh



(a) The participants performed the tasks on this interface.

(b) The continuation of their work was generated and notification was sent.

(c) They were able to refer to the generated continuation in the *proposed* condition.

Figure 18: Interface used in Study 1.

participants agreed to the research policy and were provided with a brief explanation of the usage of the Web pages in the beginning. Here, it was clarified that they would not be penalized even if they spent a long time on the tasks so that they could perform the tasks as usual. In addition, they were told that they were allowed to perform any habit they might have in performing tasks, for example, listening to music, to induce spontaneous behavior, following the procedure of Liu *et al.* [180]. They then performed three essay tasks using the Web page that corresponded to the assigned order. Every time they completed a task, they were asked to fill out a questionnaire to evaluate their experience of the task using the Web page. They were also asked to fill out another questionnaire that requested a comparison of the three conditions and additional comments after completing all the assigned tasks. The study was conducted remotely for all participants.

5.3.5 Measure

Multiple measures were prepared to examine the hypotheses. From the participants' behavior during the use of Web pages, four measures were calculated: interest retrieval time, ignorance rate, progress after resumption, and total writing time. From the questionnaire responses, cognitive load and system usability were evaluated.

Interest retrieval time In correspondence with H1, *interest retrieval time* was calculated, which denotes the time each participant spent from the moment they received a notification to the moment they recovered their interest in the task. Here, the recovery of interest was caught by finding the moment when they made any operations on the Web page (*e.g.*, moving a mouse or pressing a key) for the first time after receiving a notification. Then, the values of the *proposed* and *control* conditions were compared because the *none* condition did not present any notifications. Thus, when the interest retrieval time of the

proposed condition was shorter than that of the *control* condition, H1 is supported.

Ignorance rate Also, *ignorance rate* was calculated, which denotes a rate of interventions (*i.e.*, notifications) that were ignored by the participants. The ignorance rate would be highly correlated with the interest retrieval time (*i.e.*, a longer interest retrieval time leads to a higher ignorance rate) because the *proposed* and *control* conditions prompted a notification every T s until the participants made any operations. Still, this measure was prepared to evaluate H1 precisely. Specifically, as discussed in Section 5.1.1, there was a risk of losing the trust of the participants due to a mismatch with their expectations, fostering ignorance of the notification. Therefore, to confirm that CatAlyst can keep attracting their interest, how the ignorance rate changed as the number of interventions each participant received increased was examined. If the ignorance rate of the *proposed* condition was lower than that of the *control* condition and exhibited stable values with respect to the increase in the number of interventions, H1 is further supported.

Progress after resumption In correspondence with H2, I quantified and compared the number of characters the participants typed within a specific period after the participants recovered their interest in the task, namely, *progress after resumption*. As discussed in Section 5.2.2, they might procrastinate on the task even though they have recovered their interest. In such cases, progress after resumption would remain insignificant. Conversely, H2 is supported if the increase in the number of characters after resumption in the *proposed* condition was more significant than that in the *control* condition.

Total writing time The productivity of the participants was evaluated via two measures. The first is the *total writing time* they spent on the tasks. The average of the total writing time would be equivalent if CatAlyst did not contribute to the improvement of worker productivity since the difficulty and assignments of the tasks prepared were balanced. In other words, H3 is not supported unless the total writing time of the *proposed* condition was shorter than that of the other two conditions.

Subjective quality The *subjective quality* of the texts the participants wrote was also evaluated. This was to ensure that the reduction in the total writing time was enabled not by sacrificing the text quality, but by the improvement in their productivity. Therefore, three volunteer raters who had not participated in the study were invited to this study to rate the texts in a manner agnostic to the experimental conditions. Each text was evaluated for three aspects using a 7-point Likert scale: consistency, readability, and overall quality. Therefore, H3 is supported when the *proposed* condition shortened the total writing time while the raters' evaluation for the texts written in this condition, at least, did not deteriorate compared to the other two conditions.

Cognitive load This study also used a questionnaire to examine H4. Specifically, the six items of the NASA-TLX [107] were used to calculate the raw TLX score [37]. This allowed us to examine *cognitive load* across six factors: mental demand, physical demand, temporal demand, performance, effort, and frustration. As higher scores indicate higher cognitive load, H4 is supported when the scores of the *proposed* condition were lower than those of the other conditions.

System usability Furthermore, *system usability* was examined to complement the cognitive load in terms of evaluating the subjective experience of the participants. Here, ten

items from the System Usability Scale (SUS) [34] were adopted to compare their usability evaluation for the *proposed* and *control* conditions. The case in which the score of the *proposed* condition was higher than that of the other conditions corroborates H4 because it implies better usability of CatAlyst. Moreover, the participants were asked to rank the three conditions in order of their evaluation of system usability in the questionnaire they completed at the end of the study. Their responses were compared across the three conditions to confirm that the *proposed* condition was ranked higher than the other conditions, which leads to supporting H4.

5.3.6 Results

This study provided the following results:

Interest retrieval time Firstly, the interest retrieval time is presented in Table 5. The participants in the *proposed* condition reacted to the intervention within a much shorter period than those in the *control* condition. Welch's t-test after Levene's test confirmed that the mean of the values in the *proposed* condition was significantly shorter ($t(72.31) = 3.20$, $p = 0.002$) than that in the *control* condition, supporting H1. Furthermore, the *control* condition exhibited a greater variance in the interest retrieval time. In fact, cases where the encouraging message did not work were observed for some completely distracted participants, whose interests were recovered after more than 10 min, conforming to previous research [192].

Ignorance rate Moreover, H1 was corroborated by a comparison of ignorance rates. Corresponding to the longer interest retrieval time in the *control* condition, it presented a much larger number of notifications until the participants recovered their interest in the task, most of which were ignored, as shown in Table 6 (A). The ignorance rate of the *proposed* condition was significantly lower ($p < 0.001$) than that of the *control* condition, according to Fisher's exact test. Interestingly, no difference was found between the two conditions when only the first intervention presented to each participant was considered ($p = 0.260$). In other words, the intervention of the *control* condition attracted the participants' interest at the beginning, but they became less persuaded by the encouraging messages as they saw more encouraging messages, leading to an increase in the ignorance rate. Conversely, this suggests that CatAlyst can keep attracting the interest of workers, which aligns with the intention of introducing personalized, context-aware, and variational intervention.

In addition, this result can be explained by *perceived utility*, which Ling *et al.* [174] mentioned as a factor in maintaining the use of ML-based systems. The workers will keep interested in the intervention if they feel usefulness in the generated content (*e.g.*, as support for ideation), whereas conventional encouraging interventions do not have such utility. I later discuss this point based on the *system usability* and comments of the participants.

Progress after resumption The progress after resumption was then examined, as presented in Table 7. It was found that the participants in the *proposed* condition made more progress within T s after their interest returned to the task than those in the *control*

[6]Total number of the notifications presented to the participants as the first intervention in the *control* condition exceeds the number of participants because some of them did not recover their interest until several notifications were prompted even in the first intervention.

Table 5: Comparison of the interest retrieval time between the *proposed* and *control* conditions, which were significantly different ($p = 0.002$).

	Mean	SD
Control	135.0	± 303.7 (s)
Proposed	18.8	± 25.1 (s)

Table 6: Comparison of the ignorance rate between the *proposed* and *control* conditions. While the rate was significantly lower for the proposed condition ($p < 0.001$), no difference was observed when only the first intervention presented to each participant was considered ($p = 0.260$).[6]

	(A) **All**		(B) **First time**	
	Ignored	Worked	Ignored	Worked
Control	185	71	4	11
Proposed	5	29	0	8

Table 7: Comparison of the progress after resumption time between the *proposed* and *control* conditions, which were significantly different ($p = 0.041$).

	Mean	SD
Control	17.7	± 36.3 (chars)
Proposed	63.5	± 186.2 (chars)

Table 8: Comparison of the total writing time between the three conditions, which were not significantly different ($p = 0.174$).

	Mean	SD
Control	2557.3	± 2196.5 (s)
Proposed	1747.3	± 1660.0 (s)
None	2231.2	± 1359.8 (s)

condition, according to the t-test ($t(98) = 2.08$, $p = 0.041$). This supported H2, while implying two possibilities regarding the participants' behavior: 1) participants in the *proposed* condition could utilize the generated content to come up with what to write next, and 2) participants in the *control* condition procrastinated the task even after being guided to perform some operation on the Web page by encouraging messages.

Total writing time As shown in Table 8, no significant difference was found in the total time the participants spent between the three conditions, according to the Friedman test after Levene's test. Thus, H3 was not supported. However, we can see that the participants in the *proposed* condition spent less time on average than those in the other conditions, which implies the effect of CatAlyst. Simultaneously, the average time spent in the *control* condition was slightly longer than that in the *none* condition. This can be attributed to the fact that, while the previous work in Section 5.1.3 employed a clear criterion for detecting interruptions, such as being presented with explicit interruption cues by an experimenter [180], time-based naïve criteria was used here to examine the effectiveness of CatAlyst against false-positive detection.

Subjective quality Meanwhile, no significant difference was found in the raters' subjective evaluation of the quality of the texts the participants wrote between the three conditions, according to the Friedman test. Specifically, as shown in Table 9, all of the three aspects the raters evaluated (*i.e.*, consistency, readability, and overall quality) did not exhibit a significant difference ($p = 0.354$, 0.221, and 0.972, respectively). This result implied that CatAlyst did not deteriorate the quality of the texts the participants wrote, while H3 was not supported.

Cognitive load To evaluate H4, the participants' evaluations of their cognitive load were examined, as shown in Figure 19. One-way repeated measures ANOVA suggested significant differences in their cognitive load with respect to the frustration factor ($F(2, 22) =$

Table 9: Comparison of the subjective quality of the texts between the three conditions, which were not significantly different regarding consistency, readability, and overall quality ($p = 0.354$, 0.221, and 0.972, respectively).

	Consistency	Readability	Overall quality
Control	5.50 ± 1.21	5.25 ± 1.32	5.11 ± 1.37
Proposed	5.03 ± 1.61	5.17 ± 1.63	4.89 ± 1.65
None	5.47 ± 1.54	4.97 ± 1.46	5.06 ± 1.43

Table 10: Comparison of the participants' evaluations of the system usability between the *proposed* and *control* conditions, which were not significantly different ($p = 0.082$).

	Mean	SD
Control	62.1	± 22.8
Proposed	75.6	± 13.7

9.37, $p = 0.001$). Then, it was observed that the frustration they felt during the *proposed* condition was significantly lower than those of the *control* ($t(11) = 3.36$, $p = 0.019$) and *none* ($t(11) = 3.25$, $p = 0.019$) conditions, according to the post-hoc test with Holm correction. This result suggests the plausibility of H4.

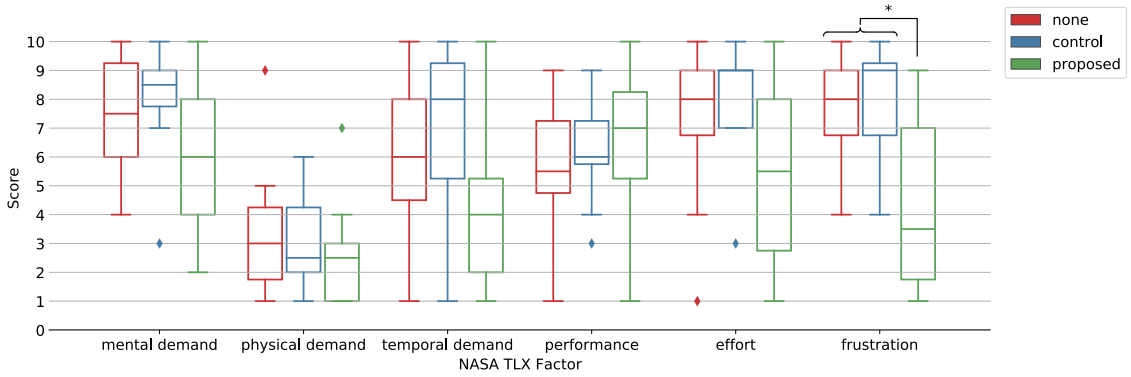


Figure 19: Participants' evaluations of their cognitive load in Study 1. The frustration they felt in the *proposed* condition was significantly lower than those of the *control* ($p = 0.019$) and *none* ($p = 0.019$) conditions.

System usability Finally, the participants' evaluations of the system usability were examined. As presented in Table 10, the scores obtained using SUS implied that the participants preferred the *proposed* condition over the *control* condition, while their difference was not significant according to the paired t-test ($t(11) = 1.91$, $p = 0.082$). At the same time, the participants' rankings regarding usability suggested that the three conditions made a significant difference in their rank order, according to the Friedman test ($\chi^2(2) = 8.91$, $p = 0.012$). Specifically, the post-hoc Nemenyi test [210] revealed that the *proposed* condition was likely to be ranked significantly higher (*i.e.*, regarded as having better usability) than the *none* condition ($p = 0.022$). In particular, eight participants responded that the *proposed* condition was the best. These results indicated that CatAlyst was favorably accepted by participants, corroborating H4. This also implied that the *proposed* condition contributed to the perceived utility of the participants, and thus, its effectiveness did not deteriorate after repeated exposure, as mentioned above.

5.3.7 User Comments

From the comments the participants left, it was confirmed that CatAlyst functioned as expected, as follows.

> Not only did the notification make me realize that I was distracted, but it also helped me get back to the task as I wondered what the AI had written.
>
> I often lost concentration and stopped working because I could not come up with the next sentence, so it was easier to get back to work with being presented with the AI's text.

Furthermore, the comment suggested that the generated continuation helped their work by providing ideas and relevant information.

> I was very surprised the first time when the text from the AI was presented. It actually fit the topic, and I was able to use that as a reference for my writing.
>
> Although I did not find consistency or uniformity in the AI's output, I found it useful for imitating its writing style and obtaining ideas for paraphrasing.

The latter comment suggested that the participant positively perceived the use of CatAlyst even though the performance of the off-the-shelf generative model was not completely satisfactory. Importantly, such a gap in performance expectations can lead to devastating trust, as discussed in Section 5.1.1. However, the presented interaction design reframes the main goal of CatAlyst to guide participants' interest instead of directly contributing to their tasks. We can infer that this induced the participants' tolerance of the AI's performance and the enhancement of their subjective usability.

The observations are further supported by examining their comments on the *control* condition, as follows:

> At first, it was a good trigger for recognizing that I was not concentrating. But, when I received the notification from the system while I interrupted the task to write a quick reply to an incoming message, I felt a little annoyed.

This comment not only emphasizes the importance of designing an intervention that also works with false-positive detection but also suggests the reason why the intervention of the control condition tended to be ignored as the number of interventions increased (Table 6). Note that one participant requested that the frequency of the notification be controllable. This conforms to the previous study that emphasized the importance of providing global control in ML-based systems [6], and thus, it was reflected in the prototype in the following long-term study.

5.4 Study 2: Writing Task in the Wild

Thus far, the effectiveness of the presented interaction design in the writing task has been validated by confirming H1, H2, and H4 through the controlled user study. Still, the results might be attributed to the fact that it might be the first time that the participants interacted with the large generative models. In other words, it was desirable to conduct a long-term study to deny the possibility that the effectiveness of CatAlyst depends on freshness. Therefore, the effectiveness of CatAlyst was evaluated in a study with a longer duration and participants having diverse motivations and purposes for writing. Ten participants used the prototype for five days and provided their perspectives and feedback on the design of CatAlyst.

Table 11: Backgrounds of the participants in Study 2.

	Gender	Writing habit	Main usage
P1	M	3 years	Novel
P2	F	8 years	Web article / Personal blog
P3	F	2 years	Diary
P4	M	0.5 years	Personal blog
P5	F	4 years	Web article / Personal blog
P6	F	rarely	University report
P7	M	10 years	Diary
P8	M	1 year	Fantasy / Lyric writing
P9	M	rarely	Private letter
P10	M	6 years	Novel

5.4.1 Implementation

The prototype used was basically not modified from Section 5.3, except that it was enabled to adjust the time threshold T based on the feedback gained in the previous study, as it would be helpful, particularly when the prototype is used in the wild. Furthermore, it was expected that participants would leave the Web page for a substantial amount of time because of other tasks in their daily lives (*e.g.*, having lunch) in contrast to the controlled setting in the first study. Thus, a feature of sending participants an email with generated content 5 minutes after they left the Web page was added based on the expectation that this feature would help participants resume the original task when they finish other tasks. Note that the content written by participants was neither accessible to experimenters nor stored on the server to preserve their privacy. It was only used for the inference by the off-the-shelf GPT-3 model and immediately discarded afterward.

5.4.2 Procedure

In the same manner as in Study 1, ten participants (P1–P10) were recruited. Table 11 summarizes their backgrounds. They self-reported their writing experience regarding the period they regularly had writing tasks. In this study, the participants first agreed to the research policy and familiarized themselves with how to use the prototype. They were instructed to use the prototype for writing freely during the study period whenever they would like, without setting any quota. After a participant spent five days, they were asked for their comments to understand how they perceived the use of CatAlyst in terms of its usability and influence on their behavior. The participants were asked a series of questions: "*what kind of writing did you use the system for?*", "*can you tell us your overall impressions of using the system?*", "*did you feel any change with and without the system?*", "*what is your impression about the AI which generated content?*", "*what was this system like for you?*", "*what would you like to see with this system in the future?*", and "*do you want to continue using the system?*". Note that the study was conducted remotely for all participants.

After the above procedure, the participants' use of CatAlyst was also examined from the interaction log. This aimed to confirm the long-term effectiveness of CatAlyst in a manner disassociated from the possible effect of the freshness in interacting with large generative models. Specifically, the *interest retrieval time* was calculated in the same manner as in

Table 12: Comparison of the interest retrieval time over the five days of the participants' use, which were not significantly different ($p = 0.791$).

Day	#1	#2	#3	#4	#5
# of detected interruptions	31	49	10	13	14
Average retrieval time	95.4	117.1	121.1	105.2	93.4
Standard deviation	81.9	91.4	102.7	95.4	89.6

the previous study to analyze its trend over the five days of their use.

5.4.3 Usage results

As presented in Table 12, no significant difference was found in the interest retrieval time across the five days, according to one-way ANOVA after Levene's test ($F\,(4, 112) = 0.425$, $p = 0.791$). This suggested that the effectiveness of CatAlyst in guiding the participants' interests did not disappear even after long exposure to it. Note that the higher values in Table 12 compared to the first study (Table 5) are attributed to the difference in the situations. Specifically, these values include the moments when the participants interrupted their writing tasks for necessary reasons in their daily lives, while the participants in Section 5.3 were asked to complete the assigned tasks as a part of the study.

In addition, Table 12 allows us to infer that the participants continued the use of CatAlyst over the five days, whereas the occasions they used can be influenced by various factors (*e.g.*, weekdays or weekends) in the uncontrolled setting of this study. Note that, since the content written by the participants was not stored, the progress they made cannot be quantified. Therefore, I then analyzed their responses in the semi-structured interviews.

5.4.4 Interview Results

Overall, participants responded positively to the question, "*can you tell us your overall impressions of using the system?*", and mentioned various benefits of the prototype. Moreover, eight (all participants except P7 and P10) showed their motivation for continued use toward the question: "*do you want to continue using the system?*" This result corroborated the support of H4, that is, the use of CatAlyst offers a favorable experience, and workers feel its usefulness. Moreover, four of them (P4, P5, P6, and P9) mentioned the novelty of the outputs as a key reason for their willingness to use the prototype.

> I've always been curious about what's generated. They were interesting. The excitement never stopped during the use and I'd not get bored. (P6)

These comments implied that the participants' favorable responses in Section 5.3 are not merely due to its freshness but thanks to the design of CatAlyst. Specifically, the variety in the outputs of the generative models plays a key role in keeping attracting workers' interest in the intervention, as discussed in Section 5.2.2. The implications obtained from the responses to the other questions are summarized below.

Effects on behavior Seven participants (P2, P3, P4, P5, P6, P8, and P9) mentioned a significant change in their behavior owing to the system as an answer to the question: "*did you feel any change with and without the system?*"

> The notifications helped me cut out the time I spent touching my smartphone. I also sometimes started or proceeded to write sentences because I was curious to see what kind of sentences the AI would generate. I found this to be effective in the initial stages, when the writing was the heaviest. (P3)

The comments were parallel to those in the short study (Section 5.3.7) and demonstrated how CatAlyst affects workers and prevents their task procrastination. In particular, P3's comment implied a *catalyst*-like effect by lowering the hurdle for resuming the task, corroborating the effectiveness of the presented interaction design.

Feelings about the accuracy of the model Next, the participants responded similarly to the question: "*what is your impression about the AI which generated content?*" More specifically, they pointed out the variance in the quality of the generated content.

> For my creative writing, the tool provided me with a good stimulus and motivation to continue writing because it allowed me to choose plots and words that I could not have reached on my own. However, I could not directly leverage it well for articulating my thought, like in a diary. (P8)

> I felt that AI has its strengths and weaknesses depending on the field of articles I write. As for abstract writing for hobbies, such as travelogues, novels, diaries, *etc.*, the AI helped me think about how to phrase words and what to do next, and it worked beyond my expectations. However, AI generation was not appropriate when it requires a great deal of checking, like using it for my job [Web writing]. (P2)

> I was trying to write a full-length novel game scenario, and there was a constraint that each character setting was already fixed. While the outputs were interesting and I had fun seeing them, I needed to manually edit them, such as wording in generated utterances. (P10)

This result highlighted the difficulty of expecting an off-the-shelf generative model to perform well in all domains without any tuning, as discussed in Section 5.1.1. However, it is noteworthy that the participants affirmatively perceived their experiences of using CatAlyst owing to its design of not aiming to directly replace part of their tasks in the first place.

Role of CatAlyst The participants' responses shed light on the three different roles of CatAlyst regarding the question: "*what was this system like for you?*"

CatAlyst as a reminder Two participants (P2 and P9) appreciated the effective reminder function:

> I used the system for my job. Due to the working-from-home situation, I often get distracted. I appreciate the reminding function as well as its generated content, which lowers the hurdle for writing. (P2)

CatAlyst as an ideator The benefit of obtaining ideas from CatAlyst was also emphasized by some participants (P1, P4, P8, and P10), who mostly performed creative writing, as follows:

> The suggested wording and story flow can be enough hints. For example, I got inspired to include a monologue of the protagonist or conversation to improve

> the tempo of the story. I used them to accelerate my writing. (P1)

This benefit is aligned with previous works supporting writing with off-the-shelf generative models and confirms the validity of our expectation that CatAlyst can induce task resumption in workers by improving their ability to perform the task (Section 5.2.2).

CatAlyst as a peer Other participants (P2, P3, P5, and P6) perceived CatAlyst as a peer or collaborator who motivated them:

> I found it like a co-worker working with me. I became curious about what they produce while I was taking a break. (P5)

> It is easier to expand my thinking by reviewing AI's outputs rather than contemplating from scratch by myself. In that sense, the AI was a peer to whom I can talk. (P3)

Their responses corresponded to our expectation that CatAlyst can gain workers' interest by stimulating their curiosity and can induce their task resumption by increasing their motivation (Section 5.2.2).

Room for further improvement of CatAlyst The participants also mentioned their requests for the system when answering the question: "*what would you like to see with this system in the future?*" Some participants (P1, P4, and P7) mentioned the need to control notification timing as follows:

> Ideally, I want to turn off the notification when I'm deeply thinking about the story flow, even if I'm not typing any words. (P1)

This point can also be mitigated by improving the detection mechanism of the moments to intervene in, considering that the current approach in the prototype is a practical but naïve method.

> I hope the system cooperates with a smartwatch because I think it would be useful to receive notifications when I am walking or something like that, so I can think about it. (P4)

This comment implied a potential need for workers to actively set an intervention timing for CatAlyst as a reminder while performing different tasks (*e.g.*, walking). Moreover, there is an interesting conflict in the comments.

> I could not wait for the AI's generation. It would have been even better if there had been a button for actively using it or more suggestions from the system. (P7)

> I like its current design. If there had been an auto-generate button, it would have been a bad experience if the output had been inaccurate. In the current case, by thinking of it as an added feature of a reminder, my expectations were relaxed and I was glad to see it for the fun of it. (P6)

While implementing such a button for actively using large generative models is feasible, we need to acknowledge that workers' trust in the system would be affected by how much they expect in the system and how accurate the provided output is, as Yin *et al.* [359] suggested. In this sense, placing a button for actively using models can lead to workers' raised expectations, as they would rely on the feature to boost their work, and thus,

their trust could be degraded if the output quality is insufficient. This point emphasizes the importance of finding an appropriate subproblem, as suggested by the strategies in Section 3.2, rather than attempting to address all visible user needs, especially when using off-the-shelf ML models.

Three participants (P1, P8, and P10) wanted to see multiple generations simultaneously.

> I would appreciate it if it could suggest several patterns of sentences at the same time so that I can do a lot of thinking when I receive the notification. (P1)

This demand can be explained by prior research suggesting that presenting multiple perspectives, including AI's suggestions, can foster one's reflection by offering discussion grounds [10]. Additionally, two participants (P6 and P10) requested CatAlyst implemented in existing editors.

> It's interesting, but I'd prefer to keep using my current editor [Google Docs] simply because it has more useful functions. If there is an API for integrating this system into my editor, I'd appreciate it. (P10)

In fact, CatAlyst can be easily integrated into online editors, as the next study shows a prototype in the form of a Chrome extension for supporting slide-editing on Google Slides.

5.5 Study 3: Slide-Editing Task

The short- and long-term studies in Sections 5.3 and 5.4 demonstrated the effectiveness of CatAlyst in writing tasks. Another study was then conducted in which participants performed slide-editing tasks with the support of CatAlyst to examine the domain extensibility of the presented interaction design. I again evaluated the hypotheses in Section 5.2.3 in a similar manner to Section 5.3 using a within-participant design with the *proposed*, *control*, and *none* conditions.

5.5.1 Task

For this study, three tasks were prepared in which participants were asked to create a slide of at least 15 pages about an aging society, online abuse, or global warming. Here, a six-page draft slide for each topic was prepared and provided to the participants to start the corresponding task. This aimed to balance the difficulty of the tasks because, without the draft, the quality of the slides and the time required to edit them would largely depend on each participant. It took roughly 30 minutes to complete a single task, and the three tasks were randomly assigned to three conditions for each participant.

5.5.2 Implementation

In this study, participants were asked to use Google Slides[7] to edit the slides after installing a Chrome extension that I developed. The extension detects the moment when the participants become distracted according to their interaction log, in the same manner as in Section 5.3.3. In the *proposed* condition, the ID of the edited slide was sent to a remote server where its continuation was generated by transforming the slide in progress into a text input to the model. Here, the same GPT-3 model as Section 5.3.3 was used without

[7] https://www.google.com/slides/about

Table 13: Comparison of the interest retrieval time between the *proposed* and *control* conditions, which were significantly different ($p = 0.002$).

	Mean SD
Control	49.5 ± 81.8 (s)
Proposed	19.2 ± 44.0 (s)

Table 14: Comparison of the ignorance rate between the *proposed* and *control* conditions. While the rate was significantly lower for the proposed condition ($p < 0.001$), the difference was not observed when only the first intervention presented to each participant was considered ($p = 0.712$).

	(A) **All**		(B) **First time**	
	Ignored	Worked	Ignored	Worked
Control	85	104	7	11
Proposed	19	70	4	11

performing any tuning for slide-editing. In addition, part of the model output was provided to a diffusion model [250] to generate an image that complements the generated slide. The slide was then appended to the original slide via the Google Slides API,[8] and the participants were prompted with a notification displaying the title of the generated slide. In the *control* condition, the same six encouraging messages as in Section 5.3.3 were used to show a notification displaying one of them at the time of distraction.

5.5.3 Procedure

This study followed the procedure of Study 1, involving twelve participants (eight males and four females; 20–59 years old[9]). Similarly, each participant completed four questionnaires: three at the end of each task and one at the end of all tasks. From their behavior during slide-editing and their responses to the questionnaires, the same measures as in Section 5.3.5 were prepared, except for the modification of *subjective evaluation* and the removal of *progress after resumption*. For the subjective evaluation, the raters were asked to evaluate each slide with respect to consistency, visual appearance, and overall quality because readability is not applicable to this task. The progress after resumption is omitted because, at the interface of Google Slide, a precise edit history that allows us to quantify the progress could not be obtained either by the extension or by the API. Therefore, considering that H2 is supported in Section 5.3.6, H1, H3, and H4 were examined using the other measures.

5.5.4 Results

This study provided the following results:

Interest retrieval time Regarding H1, similar results as in Section 5.3.6 were obtained. Specifically, the interest retrieval time was significantly shorter ($t(165.05) = 3.16$, $p = 0.002$) for the *proposed* condition than for the *control* condition, as presented in Table 13.

Ignorance rate Moreover, as shown in Table 14, the ignorance rate was also significantly lower ($p < 0.001$). It was also confirmed that, at the first intervention, the *control* condition was considerably effective compared to the *proposed* condition, while repeated exposure to encouraging messages degraded their effectiveness.

Total editing time For H3, a significant difference was observed in the time the participants spent between the three conditions, according to the Friedman test after

[8] https://developers.google.com/slides/api

[9] Note that one participant preferred not to disclose their age.

Table 15: Comparison of the subjective quality of the slides between the three conditions, which were not significantly different regarding consistency, visual appearance, and overall quality ($p = 0.311$, 0.960, and 0.475, respectively).

	Consistency	Visual appearance	Overall quality
Control	5.28 ± 1.09	5.00 ± 1.15	4.92 ± 1.16
Proposed	5.22 ± 1.02	5.03 ± 1.18	5.08 ± 1.11
None	5.42 ± 0.87	4.89 ± 1.28	5.00 ± 1.20

Table 16: Comparison of the total editing time between the three conditions. The time spent by the participants in the *proposed* condition was significantly shorter than that in the *none* condition ($p = 0.038$).

	Mean SD
Control	1776.1 ± 960.4 (s)
Proposed	1356.5 ± 1048.7 (s)
None	2794.5 ± 2866.5 (s)

Table 17: Comparison of the participants' evaluations of the system usability between the *proposed* and *control* conditions, which were not significantly different ($p = 0.101$).

	Mean SD
Control	50.6 ± 19.4
Proposed	68.8 ± 20.1

Levene's test ($\chi^2(2) = 6.50$, $p = 0.039$), as shown in Table 16. The post-hoc Nemenyi test [210] was then conducted and revealed that the time spent in the *proposed* condition was significantly shorter than that in the *none* condition ($p = 0.038$).

Subjective quality On the other hand, no significant difference was observed in the raters' evaluation of the quality of the slides the participants edited (Table 15), according to the Friedman test ($p = 0.311$, 0.960, and 0.475 for consistency, visual appearance, and overall quality, respectively). Thus, it was implied that the *proposed* condition reduced the total time the participants spent while maintaining the quality of the edited slides. This suggests that CatAlyst can contribute to improving worker productivity, supporting H3.

Cognitive load Their evaluations of the cognitive load were examined based on the participants' responses to the questionnaire. As illustrated in Figure 20, significant differences were found in cognitive load in terms of temporal demand ($F(2, 22) = 3.99$, $p = 0.033$), performance ($F(2, 22) = 3.79$, $p = 0.038$), effort ($F(2, 22) = 8.63$, $p = 0.001$), and frustration ($F(2, 22) = 7.80$, $p = 0.003$) according to ANOVA. The post-hoc test with Holm correction revealed that the scores of the effort and frustration factors in the *proposed* condition were significantly lower than those in the *control* ($t(11) = 3.60$, $p = 0.013$ and $t(11) = 3.04$, $p = 0.033$, respectively) and *none* ($t(11) = 2.79$, $p = 0.035$ and $t(11) = 2.9$, $p = 0.033$, respectively) conditions.

These results implied the advantage of the presented interaction design. That is, the design of guiding the participants' interest using the generated continuation of the interrupted work, which is context-aware and variational, would ease their frustration. This would reduce the effort required to resume the task as well. Simultaneously, as mentioned in Section 5.3.7 and also later in Section 5.5.5, the generated continuation sometimes supported their progress by providing some ideas or reference information. This corroborated the observation that CatAlyst contributes to participants' perceived utility.

System usability The participants' responses to items that measured system usability

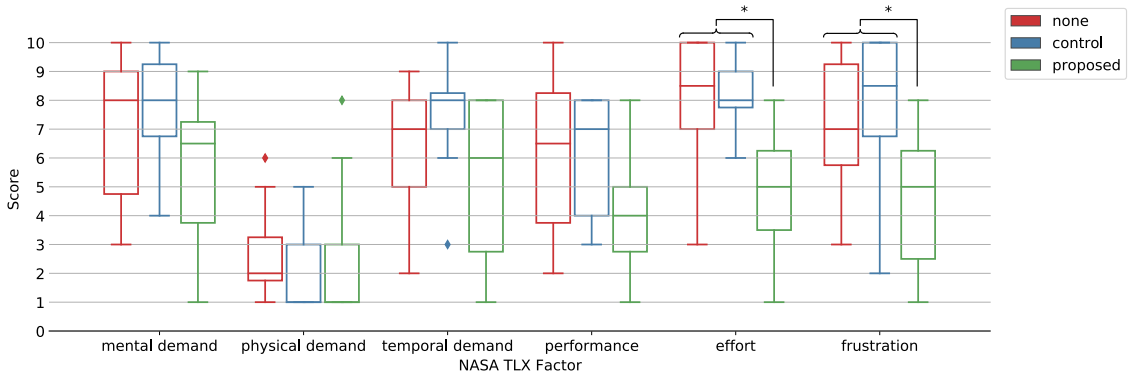


Figure 20: Participants' evaluations of their cognitive load in Study 3. The effort they made and the frustration they felt in the *proposed* condition was significantly lower than those of the *control* ($p = 0.013$ and 0.033, respectively) and *none* ($p = 0.035$ and 0.033, respectively) conditions.

further supported this observation. As presented in Table 17, the participants provided a higher SUS score for the *proposed* condition than for the *control* condition on average, although their difference was insignificant ($t(11) = 1.79$, $p = 0.101$). Furthermore, the three conditions exhibited a significant difference in the participants' rankings regarding usability, according to the Friedman test ($\chi^2(2) = 7.09$, $p = 0.029$). The *proposed* condition was likely to be ranked significantly higher than the *control* condition ($p = 0.038$), as eight participants regarded the *proposed* condition as the best. This result supported H4, and together with the above results, we can conclude that the effectiveness and extensibility of CatAlyst were confirmed.

5.5.5 User Comments

The comments of the participants highlighted the advantage of CatAlyst in drawing their interest and lowering the hurdle for resumption in the slide-editing task as well.

> I was surprised at the high quality of the automatically generated slides. I used the presented images as is and the provided text with a little editing.

> When I was notified with just a message, I was a little uncomfortable because I felt like being monitored. Regarding the notification from the AI, I was able to regard it as if the AI tried to give me an idea, so it was more satisfactory.

These comments are consistent with their evaluation of cognitive load. However, some participants pointed out the limitation of the generated continuation, which might reflect the fact that the model used was not specifically tuned for slide-editing.

> As the quality of the AI's output was not always perfect, the process of deleting the automatically added slides or extracting useful parts from the slides and merging them into my own slides was somewhat laborious and demanding.

Still, the participant commented that "it was very helpful that the AI provided content ideas, which allowed me to focus consistently on the task without wondering." In other words, it was implied that off-the-shelf generative models that are not tuned for each individual task and might be imperfect can still be useful in helping workers avoid procrastination.

5.6 Limitations

Thus far, I have demonstrated the effectiveness and domain extensibility of CatAlyst by verifying our hypotheses through two different tasks (writing and slide-editing). Specifically, the results demonstrated CatAlyst can help workers avoid task procrastination with less cognitive load through interventions made by off-the-shelf generative models. This adds an answer to RQ1 in Section 3.4, as Strategy #1 in Section 3.2 can induce interaction design that addresses the need of intellectual workers. Furthermore, the fact that its effectiveness was confirmed through a series of user studies corroborates RQ2 affirmatively. At the same time, the studies have several limitations. First, the quantitative results presented in this chapter are limited to laboratory studies in both tasks, similar to prior work [180]. This is due to the difficulty in recording the measures used in a study conducted in the wild and making a fair comparison. For example, each worker would have tasks with different difficulties and priorities (*e.g.*, deadline, interruption, *etc.*). Therefore, in the in-the-wild study (Section 5.4), our focus was on qualitatively confirming the usefulness of CatAlyst. It is desirable to conduct a longitudinal study measuring the digital well-being of workers as an outcome, as Howe and Menges did [123], and quantifying the effectiveness of CatAlyst for future work.

The mechanism of detecting workers' distractions can be further improved based on participants' feedback. The participants noticed that interventions sometimes occurred, even when they focused on the task because the naïve method in the prototypes detected the status of workers based on the interaction log. Such methods are unable to consider moments when workers contemplate deeply (*e.g.*, thinking about a plot when writing a novel) without making any interactions on the Web page. If we could implement software to be installed on their PCs, instead of a prototype Web page, it would be possible to monitor more fine-grained activities, not just typing or not on a specific Web page. Such information will help detect moments when workers are actually distracted with higher precision [298, 347, 346], which would also contribute to quantifying the effectiveness of CatAlyst by contextualizing the moment in the longitudinal in-the-wild study mentioned above. Still, I want to emphasize that CatAlyst did not degrade workers' experiences compared to conventional encouraging messages through false-positive interventions. It should also be noted that the use of large generative models poses a risk of bringing about some bias and ethical considerations, as explained in Section 2.3. In particular, we should acknowledge that stereotypes and negative propagation of particular social groups have been observed in their outputs [267, 23, 53]. Consequently, in CatAlyst, we cannot deny the possibility that such a bias is reflected in the final production of workers via the generated continuation.

5.7 Summary

This chapter presented a novel interaction design, CatAlyst, for supporting intellectual workers in their task engagement through an intervention that leverages off-the-shelf generative models. This design is delivered by considering human cognitive processes that have been discussed in ELM [230] and Fogg's model [82] based on Strategy #1 presented in Section 3.2. Specifically, CatAlyst generates a continuation of interrupted work to prompt workers to resume the task by effectively attracting their interest and inducing behavioral

change for task resumption. Note that this design is domain-extensible. Practitioners can develop interactive systems on top of CatAlyst using off-the-shelf generative models for another task domain, which would be a strong support to answer RQ1 of Section 3.4. Furthermore, the effectiveness and usability of CatAlyst were confirmed through a series of user studies, answering RQ2. At the same time, I acknowledge that the two examples of Chapters 4 and 5 would not be enough to the applicability of the strategies against open-ended user needs, and thus, I continued the exploration of other example cases in the following chapters.

6 Parametric transcription: An interaction design for enabling ultra-realistic style exploration in photo editing with off-the-shelf style transfer models

This chapter considers the application of off-the-shelf ML models to support photo editing. Particularly, we can expect that the application of GANs [95] can open up a new way of photo editing, as it has enabled ultra-realistic style transfer [142]. For example, they can be used to make a photo a user has taken (hereinafter referred to as an *original* photo) look like a particular professional shot (hereinafter referred to as a *reference* photo). This seems sufficient to address the user need to edit their own photos in a professional-like manner. However, such end-to-end style transfer is, in fact, not optimal when we consider the nature of our design processes. Our design processes are inherently exploratory—each is an open-ended journey starting with an under-specified goal [293]. Through iterative and nonlinear exploration, we refine our understanding of the space of possible designs and establish the precise form of the final product in an opportunistic and serendipitous manner [305, 296]. In other words, the intrinsic need of users is not just transferring the style of a reference photo but doing it in a customizable and exploring manner. This necessitates ML models for style transfer to serve a precise control such that each user can easily understand. However, this would be difficult without preparing a tailored model by overcoming the challenges of Section 2.3. Otherwise, the application of off-the-shelf models can result in impeding the user's understanding of possible design space and constraining their possibilities in photo editing.

To overcome this situation, I focus on the human cognitive process regarding our mental models in design processes, informed by Strategy #1 in Section 3.2. Particularly, I take note of *parametric design*, whose power has been investigated in the context of computer-aided design [30, 165]. In parametric design, all possible designs are controlled by a set of parameters, and users can easily explore variations by tweaking them [157], which helps users construct a mental model about the space of possible designs [165, 16]. Furthermore, it can be expected that, once the mental model is constructed, users can efficiently yield other variations and extend the understanding of the space, which is one of the keys to pursuing creativity [91].

Based on these advantages, this dissertation presents a novel interaction design, *parametric transcription* of an end-to-end style transfer effect, which allows users to benefit from off-the-shelf ML models for style transfer in the paradigm of parametric design. This provides users with the best parameters to imitate the effect of the models, for example, by showing a user how to set parameters (*e.g.*, brightness, contrast, *etc.*) and which photo filters to use for the original photo. Then, the user can obtain a photo that looks like the "style-baked" image generated by transferring the style of the reference photo using off-the-shelf models but is edited in a parametric manner. This allows the user to continue further edits to explore its variations. Moreover, by using the parameters and filters of an editing tool familiar to the user, such as Instagram, this interaction design can enable users to explore variations easily using the mental model they already have. This means that, by considering the nature of human design processes, we no longer need to expect the off-the-shelf models to enable a customizable and explorable style transfer in a user-wise

tailored manner, as long as they can find parameters that imitate style transfer effects.

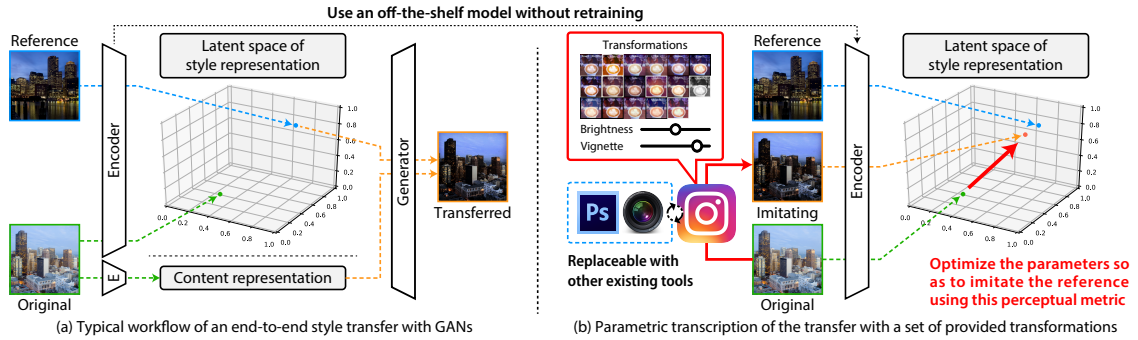


Figure 21: (a) Deep style transfer produces high-fidelity results but would be hard to utilize for exploratory design as it is performed only in an end-to-end manner. (b) The presented interaction design transcribes the style transfer effect into a set of parametric transformations available in a tool the user is familiar with (*e.g.*, Instagram), which encourages further exploration.

Still, it is not straightforward to find such parameters using off-the-shelf ML models. In particular, it would be desirable not to rely only on a specific editing tool to effectively support the exploratory processes of diverse users, considering that their familiar tools can be varied. Therefore, in this chapter, I also present a computational framework for achieving parametric transcription in a tool-agnostic manner (Figure 21). Its key components are a *perceptual metric* retrieved from an off-the-shelf model and *black-box optimization*. For the perceptual metric, this framework uses the latent representation of content-invariant styles already acquired by the off-the-shelf style transfer model. By minimizing the distance between the latent representations of the transformed result and the reference sample, it finds the optimal parameters of the provided transformations (*e.g.*, brightness, contrast, *etc.*) for imitating the style transfer effect. This optimization problem is not likely to be differentiable, but the adoption of black-box optimization [4] makes it tractable. Furthermore, black-box optimization enables users to employ various existing tools as sources of transformations.

To examine the feasibility of the presented interaction design, experiments in two different scenarios were conducted using popular smartphone apps: transcribing photo style transfer into Instagram and facial makeup transfer into SNOW. In both scenarios, subjective evaluations confirmed its capability to produce results whose quality is comparable to those produced by humans, showing its tool-agnostic applicability. Given that this interaction design would not be enabled without the consideration of human cognitive processes (*i.e.*, design processes), this suggests the applicability of Strategy #1 presented in Section 3.2 to scenarios other than those associated with productivity support (Chapters 4 and 5), which offers the answer to RQ1 of Section 3.4. Also, even though this chapter did not involve a user study, the above discussion based on the human design processes corroborates that the parameters obtained by the framework can support the creativity of users, addressing RQ2.

6.1 Mechanism and limitations of style transfer techniques

To provide background information, this chapter first explains the mechanism of style transfer techniques using GANs. Basically, cross-domain style transfer involves two networks: an encoder to map input from one domain onto its latent representation and a generator to

map the latent representation onto an output styled after another domain [178]. Formally, assuming that $\mathcal{X}$ and $\mathcal{Y}$ are two domains with different styles and $\mathcal{Z}$ is a latent space, an encoder $E_{\mathcal{X}}: \mathcal{X} \to \mathcal{Z}$ and a generator $G_{\mathcal{X}}: \mathcal{Z} \to \mathcal{X}$ are trained so that the transferred result of an original input $\boldsymbol{x} \in \mathcal{X}$ is obtained as $\boldsymbol{y} = G_{\mathcal{Y}}(E_{\mathcal{X}}(\boldsymbol{x}))$, which is expected to hold the style of $\mathcal{Y}$. These networks are trained to preserve the cycle-consistency such that $\boldsymbol{x} \approx G_{\mathcal{X}}\left(E_{\mathcal{X}}\left(\boldsymbol{x}\right)\right)$. Thus, the latent representation $\boldsymbol{z} \in \mathcal{Z}$ is considered not to be specific to either domain but to indicate style-invariant content information.

Some methods have achieved higher quality by disentangling the content and style representations [164, 130]. For example, Lee *et al.* [164] prepared three latent spaces: $\mathcal{Z}^c$ for the content representation, and $\mathcal{Z}^s_{\mathcal{X}}$ and $\mathcal{Z}^s_{\mathcal{Y}}$ for the style representations of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Using a content encoder $E^c_{\mathcal{X}}: \mathcal{X} \to \mathcal{Z}^c$, a style encoder $E^s_{\mathcal{Y}}: \mathcal{Y} \to \mathcal{Z}^s_{\mathcal{Y}}$, and a generator $G_{\mathcal{Y}}: \mathcal{Z}^c \times \mathcal{Z}^s_{\mathcal{Y}} \to \mathcal{Y}$, we can obtain the transferred result $\boldsymbol{y} = G_{\mathcal{Y}}(E^c_{\mathcal{X}}(\boldsymbol{x}), E^s_{\mathcal{Y}}(\hat{\boldsymbol{y}}))$ that resembles the style of a specific reference sample $\hat{\boldsymbol{y}} \in \mathcal{Y}$. As described in Section 6.2.2, the presented framework basically assumes the availability of the style encoder $E^s_{\mathcal{Y}}$ for defining a perceptual metric. Note that the presented framework does not use the generator since provided parametric transformations take their role in it.

Though these methods succeeded in yielding high-fidelity results, when it comes to the use of their off-the-shelf models for design support, they would not be sufficient in terms of explorability. One possible approach to allow the exploration of the variations of the result is blending the latent representation of multiple reference samples [336]. However, exploring a blend parameter between latent representations until finding a satisfactory result can be counterintuitive for users, especially considering the nonlinearity of the latent space [159]. For example, given two reference samples $\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2 \in \mathcal{Y}$, the style of the transferred result $G_{\mathcal{Y}}(E^c_{\mathcal{X}}\left(\boldsymbol{x}\right), \alpha E^s_{\mathcal{Y}}\left(\hat{\boldsymbol{y}}_1\right) + (1-\alpha)\, E^s_{\mathcal{Y}}\left(\hat{\boldsymbol{y}}_2\right))$ can be changed gradually between a $\hat{\boldsymbol{y}}_1$-like one and a $\hat{\boldsymbol{y}}_2$-like one by tweaking their blending parameter α. However, it would not be easy for a user to find the reference samples and the blending parameter that are able to produce the desired result. In addition, while more than two reference samples may be needed to ensure satisfactory variations, the more reference samples are used, the greater the number of blending parameters required, which complicates the parameter exploration.

Another approach is modifying the architecture of GANs to allow users to specify attributes in the style to be transferred [110, 102]. However, since these methods are designed to explicitly learn the attribute information, this requires a tailored ML model trained using a dataset annotated with regard to all attributes that can be transferred, which poses the challenges of Section 2.3. Moreover, unlike the blending approach described above, the user cannot control the degree of the transfer. Therefore, to leverage off-the-shelf ML models for style transfer in exploratory design processes, I present a new interaction design in a manner informed by the strategies explained in Section 3.2. Specifically, it transcribes a transfer effect into parametric transformations available in a tool a user is familiar with.

6.2 Parametric transcription

As explained at the beginning of this chapter, the presented interaction design is intended to allow users to benefit from off-the-shelf ML models for style transfer in their exploratory design processes. Before introducing the details of the computational framework to enable

it, this section explains why its design of merging style transfer techniques with parametric design is suitable for addressing the user needs.

6.2.1 Design rationale

Let us get back to the example of photo editing to understand the importance of explorability in design processes. In this case, the user has two photos, *i.e.*, original and reference photos, and wants to exploratorily edit the original photo by using the reference photo as a loosely specified initial goal. The user may first try to make some attributes in the style of the original photo (*e.g.*, brightness or contrast) look like those of the reference photo, but then the user may want to change the attributes to modify or to change the degree of the imitation randomly. Through iterations of such unstructured trials and errors, the user refines their understanding of the space of possible designs around the original and reference photo and progressively establishes the precise form of the design goal. However, as discussed in Section 6.1, the concurrent style transfer techniques are not designed to serve such exploratory design processes. In particular, the end-to-end transfer methods would not allow exchangeable combinations of step-by-step explorations consisting of trials and errors. Moreover, they would not facilitate a user's intuitive understanding of the possible design space, as how such a transfer affects the result is not easily predictable by the user.

This motivates us to take advantage of parametric design, which has been established by researchers and practitioners to facilitate an exploratory journey based on human cognitive processes. Again, it is known to induce creativity by helping users construct and expand their mental models about the space of possible designs [165, 91]. Furthermore, the combination of the presented interaction design with existing tools can foster the exploratory design processes from the following perspectives.

Expressive parameter sets In parametric design, the parameter set inherently constrains the space of potential designs, and thus, it is important that the parameter set possesses sufficient expressiveness to cover all of the user's intentions. In this regard, we can expect that the expressiveness of existing editing tools that a specific user is familiar with would align with the user's design intentions.

Transparent transformations To facilitate users' exploratory design processes, it is essential to enable users to predict how each parameter affects the result, sometimes even without actually trying them, which is often reflected in many editing tools for parametric design. Moreover, we can expect that users have an understanding of the transformation behaviors of a tool that they are familiar with.

Non-destructive trials and errors As mentioned above, exploratory design processes comprise nonlinear and unstructured trials and errors, such as revoking a specific effect that was manipulated at the beginning of the design process. We can benefit from existing tools that enable users to perform a lot of trials and errors iteratively by reverting a specific transformation or exchanging it with other transformations.

By enabling these points with off-the-shelf ML models for style transfer, the presented interaction design can realize seamless integration of ultra-realistic style transfer techniques into exploratory design processes, as illustrated in Figure 22.

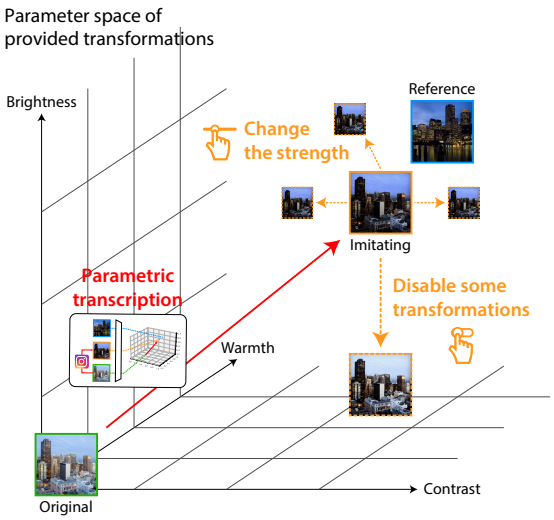


Figure 22: Since the presented interaction design shows the parameters to imitate style transfer within the provided parametric transformations, the user can continue further edits in the imitating result to explore its variations (*e.g.*, disabling some transformations or changing the strength) in the parameter space.

6.2.2 Computational framework

Such an interaction design can be realized in the following manner. Let us consider an original input $\boldsymbol{x}$ and a reference sample $\hat{\boldsymbol{y}}$ as well as a set of parametric transformations $\{T_1(\cdot;\theta_1), T_2(\cdot;\theta_2), \cdots, T_N(\cdot;\theta_N)\}$ that are controlled by parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_N\}$. N denotes the number of transformations available in an existing tool to be used for parametric transcription. We want to optimize $\boldsymbol{\theta}$ so as to make the transformed result $\boldsymbol{y} = T_1 \circ T_2 \circ \cdots \circ T_N(\boldsymbol{x};\boldsymbol{\theta})$ as similar to $\hat{\boldsymbol{y}}$ in terms of their styles as possible.[10]

This framework assumes an off-the-shelf model that can measure the perceptual distance in their styles. For example, an encoder network $E^s_{\mathcal{Y}}(\cdot)$ that outputs the latent representation of the style, such as the ones mentioned in Section 6.1, can be used here. In other words, what this framework have been calling *style* is defined by the off-the-shelf model, or indeed, the data used for its training. Now, the optimization target can be formulated as

$$\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}} \| E^s_{\mathcal{Y}}(T_1 \circ T_2 \circ \cdots \circ T_N(\boldsymbol{x};\boldsymbol{\theta})) - E^s_{\mathcal{Y}}(\hat{\boldsymbol{y}}) \|.$$

In this case, since the encoder network in GANs is trained to output latent variables that match a specific prior distribution, we can choose an appropriate norm function to measure their distance depending on it. For example, L_1-norm is empirically known to perform well for the comparison of latent variables from the prior $\mathcal{N}(0,1)$ [130].

In sum, this framework is characterized by four elements: original input, reference sample, set of parametric transformations, and perceptual metric. Its goal is to exploit the provided transformations applied to the original input so as to obtain a transformed result having a style similar to that of the reference sample. This would be easy for humans but is challenging for computers because the transformed result cannot be directly compared to the reference sample, as they have different contents. The perceptual metric is thus introduced by leveraging an off-the-shelf model, which enables the comparison between their styles and the optimization of the parameters of the transformations as formulated above.

[10] For simplicity, the composed transformation $T_1 \circ T_2 \circ \cdots \circ T_N(\cdot;\boldsymbol{\theta})$ is regarded to behave as $\mathcal{X} \to \mathcal{Y}$.

By exchanging the four elements, this framework can be applied to various situations, as presented in Sections 6.3 and 6.4.

Another key point is the use of black-box optimization algorithms, which frees this framework from constraints on the properties of parametric transformations, such as their differentiability or smoothness. At the same time, it enables users to freely replace the set of parametric transformations so that they can utilize various existing tools to meet their demands. Here, *Bayesian optimization* [264] is employed since it is efficient in terms of the number of function evaluations and allows discrete variables to be optimization targets. In the following scenarios, Optuna [4] was used owing to its ease of integration. Note that the off-the-shelf model used for the perceptual metric is not limited to the encoder network in GANs but can be based on other networks, such as variational autoencoders [71] and flow-based models [151], as long as they map the style to a latent representation well. Moreover, this independence from specific models allows us to incorporate unseen off-the-shelf ML models to be proposed in the near future for domains other than those the following scenarios consider.

6.3 Experiment 1: Photo style transfer in Instagram

This chapter first presents the scenario of transcribing photo style transfer using the presented framework. To demonstrate that it can employ various tools as a source of parametric transformations, this experiment considers transcribing into Instagram.

6.3.1 Related work

There are many methods that leverage ML technologies for photo style transfer [142]. For example, Li *et al.* [170] extended the method proposed by Luan *et al.* [183] to incorporate with an evaluation network, in an analogous manner to GAN-based methods described in Section 6.1. As discussed there, however, these end-to-end methods are not designed to facilitate the intuitive exploration of the variations of the transferred results. I acknowledge that there are some studies that predict suitable parameters, such as saturation and brightness, for photo enhancement instead of performing such an end-to-end transfer [221, 126]. Nevertheless, these methods are aimed at obtaining parameters to realize a professional-like appearance by assuming the availability of a dataset of professionally edited photos. Thus, they are not able to serve exploratory processes that use an arbitrary reference photo as a loosely specified goal.

6.3.2 Implementation

Again, this scenario considers optimizing the parameters of transformations available on Instagram. However, Instagram provides its photo editing functions in iOS and Android apps and does not provide APIs, which led to the use of UIAutomator, a testing library for Android apps, to control the Instagram app running on an Android emulator. For a perceptual metric, an off-the-shelf model (encoder network) of neural photo style transfer published by Li *et al.* [170] was used, which was trained to encode a reference photo into its style representation during the end-to-end photo style transfer. In each optimization step, the framework first sampled parameters to try next from Optuna [4]. Then, via UIAutomator, it applied all transformations, such as adjusting brightness or vignette, as

Table 18: Distribution of the collected responses indicating how many times the imitating photos in each group (obtained by the presented framework or by the human participants) were placed at each rank. Ranked first means that the evaluators judged the imitating photo to be the one most similar to the given reference photo.

Group	#1	#2	#3
Parametric transcription	105	87	108
Participant A	92	99	109
Participant B	103	114	83

Table 19: Result of the ART test to the responses in Table 18. Query refers to the pair of the original and reference photo, and its effect indicates the difference in the responses among the questions.

Effect	F	df	p-value
Query	0.02	9	0.999
Group	0.88	2	0.417
Query × Group	7.15	18	< 0.001

well as choosing a photo filter from 24 provided filters based on the sampled parameters. The transformed photo was cropped from the screenshot and provided to the off-the-shelf model to compare its style with that of the reference photo. Finally, the perceptual distance to the reference photo was returned to Optuna as a score to be minimized until 1,000 iterations were performed. Note that Instagram yields the same result for the same set of parameters without regard to the order of manipulating the parameters on the interface, which freed us from optimizing the order.

6.3.3 Subjective evaluation

Here, the imitating quality of the presented framework was evaluated in comparison with that of humans. For the same pair of original and reference photos, three imitating photos were obtained, one by the framework and the others by human participants. The imitating photos were evaluated by human evaluators recruited using Amazon Mechanical Turk (AMT) in regard to their similarity to the reference photo.

Procedure First, imitating photos were prepared using the presented framework for ten pairs of original and reference photos, which were taken from samples released by Luan *et al.* [183]. For comparison baselines, two participants who had experience in using Instagram were recruited and asked to edit the same original photo using Instagram to look like the corresponding reference photo. Figure 23 shows some examples. Note that comparison with the previous methods for photo style transfer would not be meaningful because, as described in Section 6.2.1, they perform an end-to-end transfer, whereas our aim is finding parameters for transformations available in Instagram to enable exploratory design processes.

Then, 30 evaluators were recruited in AMT and asked to fill out a questionnaire consisting of ten questions. In each question, they were presented with a reference photo and three imitating photos, one of which was obtained by the presented framework and the others by the human participants, and were instructed to rank them in order of their similarity to the reference photo. Their responses were analyzed using aligned ranked transform (ART) [259], which can analyze ranked data nonparametrically. This allows us to examine the effect of the ways to obtain the imitating photos (*i.e.*, by the presented framework or by the human participants) in a manner similar to a two-way ANOVA test.

Results The distribution of the collected responses is shown in Table 18. The ART analysis indicated no main effect from how the imitating photos were obtained, as indicated

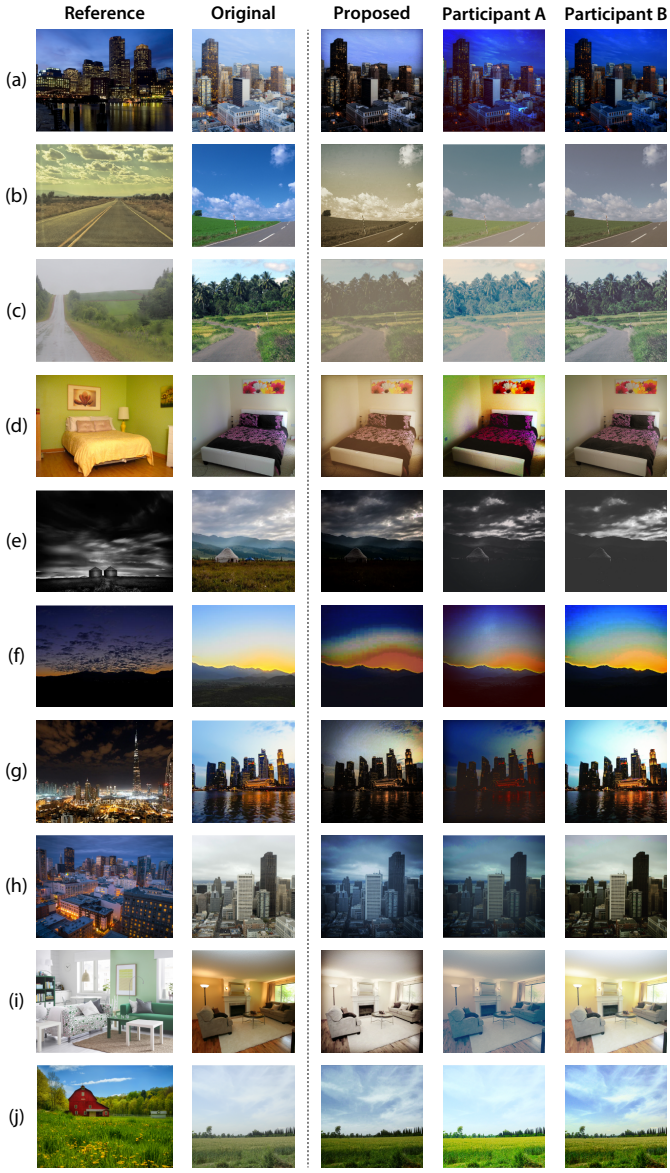


Figure 23: Examples of the original and reference photos as well as the imitating photos obtained by the presented framework and human participants. The photos in the three right columns were obtained by using Instagram to edit the original photos.

in Table 19, and thus implied that the presented framework demonstrated performance comparable to that of humans in imitating the style of the reference photo. In other words, the presented framework is applicable to the scenario of transcribing photo style transfer.

In addition, agreement among the evaluators for each question was examined by using Kendall's W. As a result, Figure 23 (d) and (e) were two of the cases that showed the lowest agreement ($W = 0.021$ and 0.048, respectively). Looking at the imitating photos for these cases, the dispersion of the responses for Figure 23 (e) can be attributable to the almost identical appearance. On the other hand, the imitating photos in Figure 23 (d) evoked the dispersion, though they seemed to vary in appearance. We can infer that this was caused by the discrepancy between the original and reference photos, which made it difficult to imitate the style using Instagram, and thus, none of the imitating photos was similar enough to obtain the consensus of the evaluators. This dispersion would be resolved by using editing tools with more expressiveness than Instagram.

Table 20: Distribution of the collected responses indicating how many times the imitating selfies in each group (obtained by the presented framework or by the human participants) were placed at each rank.

Group	#1	#2	#3
Parametric transcription	118	92	90
Participant A	113	113	74
Participant B	69	95	136

Table 21: Result of the ART test to the responses in Table 20. Unlike Table 19, it suggests that how the imitating selfies were obtained has a significant effect.

Effect	F	df	p-value
Query	0.05	9	0.999
Group	18.97	2	< 0.001
Query $\times$ Group	4.73	18	< 0.001

6.4 Experiment 2: Facial makeup transfer in SNOW

Next, I considered transcribing facial makeup transfer into SNOW using the presented framework. SNOW is one of the popular iOS and Android apps providing functions for editing selfies with various makeup features.

6.4.1 Related work

As with photo style transfer, facial makeup transfer has involved various ML techniques, such as GANs [100] or flow-based models [46]. For example, Gu *et al.* [100] exploited disentangled latent representations of a face content and a makeup style, as described in Section 6.1. However, these methods aimed to obtain a transferred selfie but did not allow users to explore the variations of the results, which motivated us to apply parametric transcription.

6.4.2 Implementation

Since SNOW also does not provide APIs, the Android emulator was used with UIAutomator, as in Section 6.3.2. For a perceptual metric, an off-the-shelf model (encoder network) of neural facial makeup transfer published by Gu *et al.* [100] was used, which was trained to encode a reference selfie with makeup into the latent representation of its makeup style. Here, although the original end-to-end facial makeup transfer had a generator to produce a facial image from this latent representation, only the encoder was used. Then, Optuna was used for 1,000 iterations to find parameters of the makeup transformations, such as lip color and eyebrows, that minimize the distance to the reference selfie in the latent space.

6.4.3 Subjective evaluation

Procedure A subjective evaluation was conducted in the same manner as described in Section 6.3.3. Ten pairs of the original and reference selfies were randomly selected from the dataset of Gu *et al.* [100] to prepare imitating selfies by using the framework. Two participants were also recruited and asked to edit the original selfies to look like the reference selfies by using SNOW. Figure 24 shows the examples. Then, 30 evaluators were recruited in AMT and asked to rank the imitating selfies ten times in the same manner as before. Their responses were later analyzed using ART [259].

Results The distribution of the collected responses is shown in Table 20. Different from Section 6.3.3, the main effect of how the imitating selfies were obtained was observed

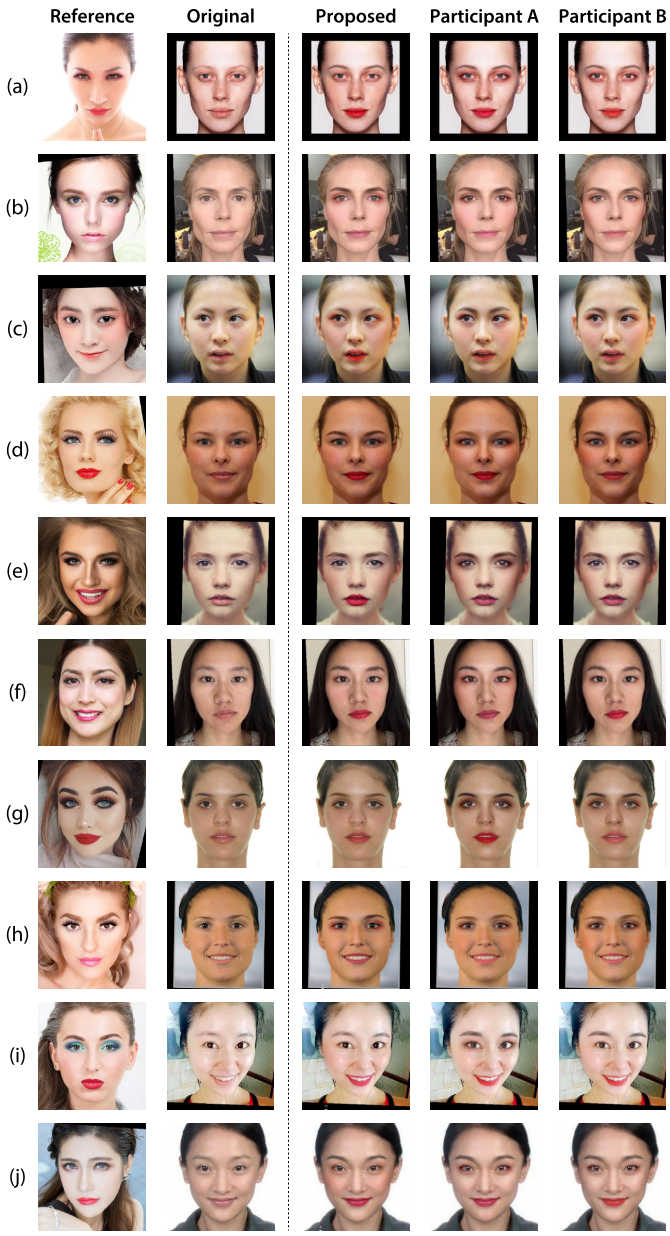


Figure 24: Examples of the original and reference selfies and the imitating selfies obtained by the presented framework and the participants.

($p < 0.001$), as presented in Table 21. Thus, Tukey's post-hoc analysis was conducted, which revealed significant differences between the selfies edited by Participant B and those either edited by Participant A or obtained by the presented framework ($p < 0.001$) but no significant difference between selfies edited by Participant A and those obtained by the presented framework ($p = 0.57$). Therefore, given Table 20, the presented framework outperformed one human participant and showed quality comparable to that of the other in regard to imitating the makeup style of the reference selfie. Along with Section 6.3.3, the result suggests that this framework is applicable to various situations and can automatically obtain parameters yielding human-level quality. In other words, it can enable users to benefit from off-the-shelf ML models for style transfer in an explorable manner within existing tools by presenting the parameters.

6.5 Discussion

The two experiments have demonstrated the applicability of the presented framework in various scenarios, corroborating the feasibility of the presented interaction design. This supports Strategy #1 presented in Section 3.2, answering RQ1 and (indirectly) RQ2 of Section 3.4. To facilitate the further application of the framework, I conclude this section by discussing its distinction with related approaches and its limitations.

6.5.1 Comparison with existing optimization-based techniques

Prior to the presented framework, many optimization-based techniques have been proposed for computational design support [193, 262], given that most design processes are parameterized by a set of either continuous or discrete parameters in current editing tools (*e.g.*, using control panels including many sliders) [157]. In particular, researchers have developed human-in-the-loop optimization techniques to deal with fuzzy optimization criteria that involve perceptual metrics or preferences. For example, evolutionary computation has been leveraged in combination with an iterative human evaluation to solve perceptual optimization problems [292]. Bayesian optimization has also been used for human-in-the-loop optimization frameworks [33, 158] because it requires only a small number of function evaluations to find a good solution [264] and thus reduces the amount of human evaluation required when used in human-in-the-loop optimization. For situations where the distance to the reference sample can be calculated by computers, optimization-based methods that work without human evaluation have been employed. Such an approach for fitting parameters to replicate a reference sample via optimization is often referred to as *inverse design* [193]. Similary, inverse *procedural* modeling [294, 125] searches for an optimal combination of parametric procedures and their parameters for replicating a given reference sample. One of its primary motivations is to support further editing of the result by tweaking the parameters, which is common to our motivation.

These frameworks differ from the presented framework, however, in that it does not assume that the reference sample and replicating result can be directly compared, *e.g.*, by pixel-wise comparison. Specifically, in transcribing style transfer effects, even applying the optimal transformations to the original input does not yield a transformed result that is identical to the reference sample because the result and the sample have different contents. Thus, while inverse design tries to exactly reproduce the reference sample based on a direct comparison, this framework is required to perform the optimization based on a content-invariant comparison, as illustrated in Figure 25. The presented framework addressed this point without the help of human evaluation by leveraging an off-the-shelf ML model to obtain a perceptual metric, which enabled new application scenarios that are intractable to conventional inverse design approaches. Note that, by focusing only on the computable aspects as the optimization objective, we can deal with style transfer in a similar manner to a conventional parameter-fitting problem, like using color transfer methods based on the color distribution comparison [243, 233] rather than photo style transfer. However, the color distribution is only one aspect of the photo style, and the restriction imposed by such a computable comparison undermines the expressiveness of the obtained results.

Moreover, the tool-agnostic applicability of the framework is another of its advantages. In particular, I acknowledge some methods in inverse design that involve a perceptual

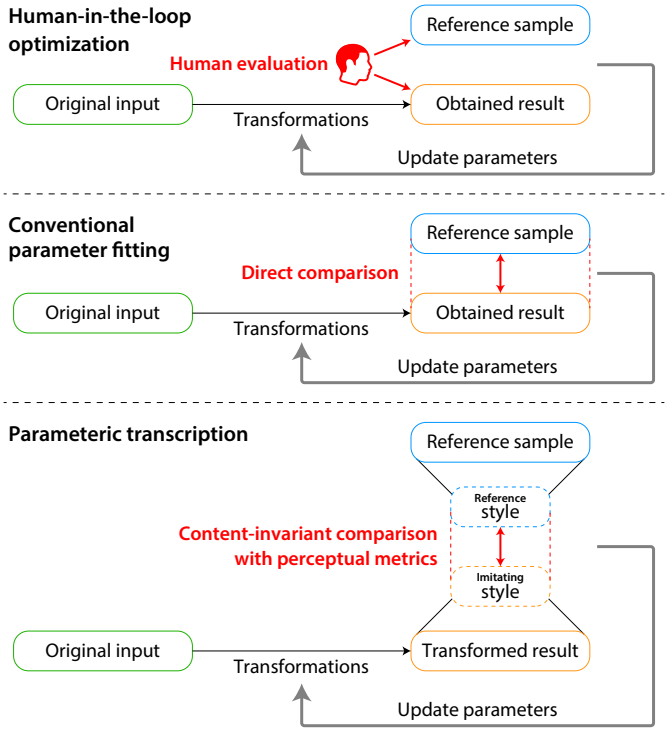


Figure 25: Difference between the presented framework and previous methods assuming human evaluation or parameter-fitting situations. Since the framework is for transcribing style transfer that is content-invariant, the transformed result cannot be directly compared with the reference sample and thus requires perceptual metrics.

metric, analogously to the presented framework. Shi *et al.* [269] optimizes a procedural texture model to resemble a reference texture, in a similar manner as Hu *et al.* [125], leveraging a perceptual metric extracted from an off-the-shelf model to perform a style-aware comparison, instead of the pixel-wise comparison. Still, this method is specifically designed for a custom-made texture model, as it relies on the differentiability of the texture model.

6.5.2 Limitations

Nevertheless, the presented framework has some technical limitations. For example, the running time can be a hurdle for its practical use. In the case of transcribing photo style transfer into Instagram, it took about an hour for 1,000 iterations. This is partially due to the black-box optimization but mainly due to the overhead caused in manipulating the Instagram app on an Android emulator using the testing library. As in humans editing photos on Instagram, it takes at least a second for a single iteration applying all parameters on the interface of the app. The running time can be shortened by performing the optimization process in parallel or by targeting applications that provide APIs (*e.g.*, Photoshop or Blender) to reduce the overhead.

The presented interaction design also has several intrinsic limitations. First, the quality of parametric transcription is constrained by the expressiveness of available parametric transformations. When the styles of the original input and the reference sample are so different that they are difficult to imitate with any combination of the available transformations, the imitating results will not be so faithful. In particular, content-aware style transfer,

such as artistic portrait stylization [357], would be challenging to imitate by parametric transformations available in existing photo editing tools. In addition, it does not cover domains for which there is no tool providing appropriate parametric transformations. For instance, there are some studies on transferring the style of human motions [363, 115]. Still, it is difficult to apply the presented framework to transcribing motion style transfer parametrically because such human motion data is usually represented by the joint positions on a frame-by-frame basis. Thus, there is no established mechanism to control it with a set of parametric transformations.

Lastly, while the power of connecting with the parametric design was discussed in Section 6.2.1, it has not entirely clarified how parametric transcription can contribute to users' creativity. For example, by taking advantage of the interaction design, it is possible to realize a new design process starting from the imitation. Specifically, the parameters provided by it can be utilized in various ways other than making the original sample similar to the reference sample, *e.g.*, making the original sample dissimilar by applying the obtained parameters inversely. Such imitation-driven explorations would provide new inspirations for creators, as Dalí [62] emphasized the importance of imitation as a source of creativity when he said that "those who do not want to imitate anything, produce nothing." Delving deeper into how the presented interaction design can be utilized would add further insights to the answer to RQ2.

6.6 Summary

This chapter introduced a new interaction design—parametric transcription of style transfer effects—and a computational framework enabling it. Given an original input and a set of parametric transformations, this framework automatically optimizes their parameters to make the transformed result have a style similar to that of a reference sample by leveraging a perceptual metric retrieved from an off-the-shelf ML model for style transfer. Then, unlike end-to-end style transfers, the interaction design offers explorability of the variations of the result by providing parameters of the transformations so that users can further edit the result intuitively in a third-party editing tool. Through two experiments, it was confirmed that the framework can be applied to various situations while achieving human-comparable quality, suggesting the feasibility of the presented interaction design to contribute to users' creativity by exploiting the intrinsic advantages of parametric design. In other words, the effectiveness of Strategy #1 presented in Section 3.2 was exhibited through the fact that they can derive interaction design to address user needs not only regarding productivity support but also for creativity support. This directly answers RQ1 and provides supporting evidence for RQ2, given the clear advantage of the integration of parametric design.

7 IteraTTA: An interactive design for guiding novice users in music composition with off-the-shelf text-to-audio models

This chapter also focuses on creativity support scenarios, specifically for novice users, given that advances in ML technologies open up novel ways for a diverse group of individuals to engage in creative processes [84, 206]. Specifically, music generation models can foster creative expression among novice users, who may not necessarily possess formal musical knowledge [40, 249]. Consequently, several approaches have been proposed to enable users to control various musical attributes of generated audios, such as specifying the note or rhythm density [295, 47] and chord progression [317, 61, 318]. Text-to-audio models [2, 177] are promising in terms of allowing users who are not familiar with the concepts of such musical attributes to generate their own sounds. This raises the expectation that providing novice users with off-the-shelf text-to-audio models will enable their creative musical expressions.

Nevertheless, I argue that such a naïve solution would not directly contribute to supporting the creativity of novice users. For example, the text-to-audio models rely on annotated labels of music clips presented in their training datasets [90, 149, 73], which primarily consist of musical descriptions such as genres, instruments, and moods. Therefore, providing such information as a text prompt is crucial for enabling fine-grained control over generated music audios. However, this may prove challenging for novice users due to disparities in artistic vocabulary among individuals with varying levels of musical knowledge [290]. Experimentally, it has been suggested that non-musicians tend to rely more on abstract concepts, such as the pleasantness or complexity of music, when appreciating musical pieces [97], which may pose difficulties in fully exploring various text prompts. This necessitates the preparation of tailored ML models attuned to the vocabulary of novice users, which poses not only the challenges mentioned in Section 2.3 but also the risk of losing fine-grained control over the generated audios.

To overcome this challenge, this chapter presents *IteraTTA*, an interaction design dedicated to the text-to-audio (TTA) music generation processes of novice users (Figure 26). This is based on Strategy #2 in Section 3.2 and designed to allow them to learn how to communicate with off-the-shelf text-to-audio models effectively. It first presents novice users with an opportunity to encounter new vocabulary by transforming their loosely-specified theme phrase into variations of relevant, rich-described text prompts. Analogously to Chapter 6, this interaction design enables iterative exploration of both text prompts and audio priors so that they can gain a comprehensive understanding of the space of possible results and how each word in the text prompts can be mapped in the space. Here, the effectiveness of this interaction design was confirmed by deploying it as a publicly available Web-based interactive system, which allowed us to analyze the diverse ways in which users utilized it in their creative processes. The results and discussions shed light on ways to utilize off-the-shelf models developed in the music information retrieval (MIR) community to unleash the creativity not only of expert users [7] but also of individuals with varying degrees of musical knowledge. They also demonstrate that Strategy #2 can generate an effective interaction design harnessing off-the-shelf models, indicating a positive response to RQ1 and RQ2 of Section 3.4.

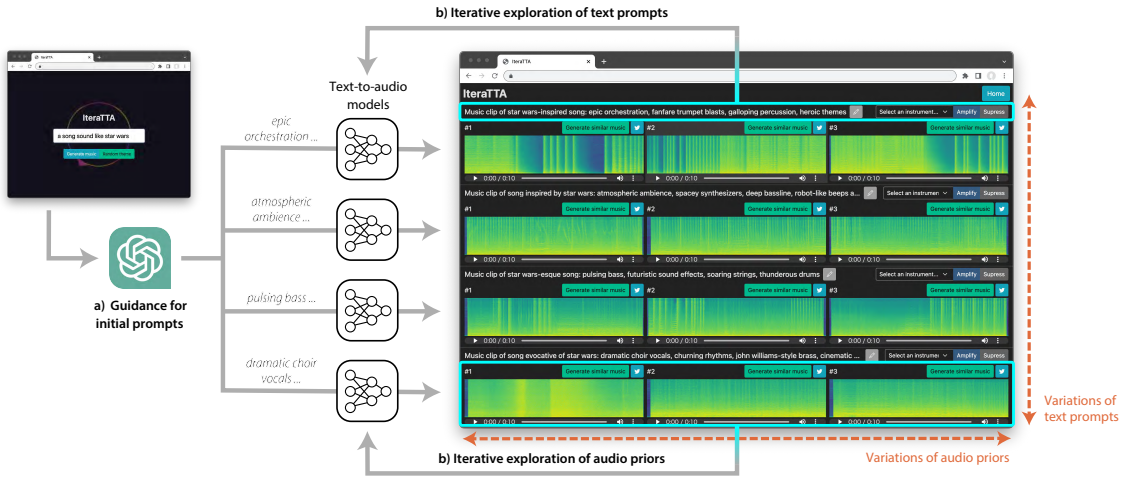


Figure 26: IteraTTA is an interaction design dedicated to allowing novice users to show their creativity in text-to-audio music generation processes. It provides a) computational guidance for constructing initial prompts and b) dual-sided iterative exploration of text prompts and audio priors.

7.1 Related work

To situate the presented interaction design, this chapter first reviews existing techniques for music generation and interaction designs for supporting them.

7.1.1 Music generation techniques

Music generation has been one of the central topics with the MIR community [246, 80, 176, 32, 67], and various ML technologies, specifically generative models, have been widely employed for this purpose [139, 67]. While methods for symbolic music generation that output MIDI files have been popular [353, 72, 129, 131, 124, 202, 360], some methods use generative models to directly output audio for leveraging their expressiveness [69, 38, 133, 226]. For example, Jukebox [69] and RAVE [38] use variational autoencoders and autoregressive models trained on large-scale music datasets to generate diverse music audios.

Controllability in music generation has been also emphasized [295, 47, 317, 61, 318, 3, 132, 212] because it is vital to open up its applications for supporting users' creative processes [315, 128]. For instance, Music FaderNets [295] allows users to modify the rhythm and note densities of generation results, while Music SketchNet [47] enables them to specify pitch contours and rhythm patterns. Wang *et al.* [317] and Dai *et al.* [61] have proposed methods to further constrain the chord progression of generation results. However, as mentioned at the beginning of this chapter, users are not always familiar with such concepts, and then, they would have difficulties in using these methods to obtain the music audios they want to generate. Note that some methods [132, 212] provide perceptual control, which does not require extensive musical knowledge, such as emotion-based musical generation. Nevertheless, they are based on Russell's valence-arousal model [256] consisting of four classes, which limits the range of controls and may hamper users' agency [111] when the methods are used to support their creative processes.

In this context, concurrent text-to-audio models [2, 177] can be an effective tool for such novice users. These models learn the relationship between music audios and their text descriptions (more specifically, latent representations encoded from the descriptions by

RoBERTa [181]) and use it to guide results in generating new audios from an inputted text (*i.e.*, text prompt). As RoBERTa is capable to encode text prompts with variable length and content, the models can provide flexible control without requiring specific musical knowledge of rhythm patterns or chord progressions. Moreover, they allow users to constrain generation results not only by text prompts but also by audio priors, ensuring that the results have similar characteristics to the priors. For example, the diffusion model [250] employed by AudioLDM [177] usually uses Gaussian noise for the seed of its generation process, but by using a noise-infused audio prior, we can obtain generation results preserving the characteristics of the provided audio.

Here, text-to-image models that use similar schemes have been shown to unleash the creativity of novice users, allowing them to iteratively explore open-ended variations of text prompts [224] and customize their intermediate results by specifying image prior constraints [86]. Similarly, text-to-audio models can be leveraged to provide users with such iterative exploration or customization. However, it is also expected that text-to-audio music generation processes may pose several specific difficulties, as explained at the beginning of this chapter. This motivated the exploration of interaction designs that can overcome these address challenges to leverage off-the-shelf ML models to address the needs of novice users.

7.1.2 Interaction design for music generation

There is a series of research on building interactive systems to let users use music generation techniques effectively [274, 127, 182, 242, 370, 373, 374]. MySong [274], for instance, involves a music accompaniment generation model, with which users can control the happiness or jazziness of generation results. Louie *et al.* [182] proposed an interactive system for novice users so that they can use a symbolic music generation technique with control of happiness or randomness. This also allows users to constrain generation results by providing music priors, which was experimentally confirmed to be effective in iteratively refining the results. Zhou *et al.* [373, 374] utilized a user-in-the-loop Bayesian optimization technique to enable novice users to explore melodies composed by a generative model iteratively. These interaction designs underscore the significance of providing controls and supporting iterative exploration to facilitate the creativity of novice users using music generation techniques. Consequently, providing the off-the-shelf text-to-audio models to novice users would be highly suitable for this purpose, as they offer more flexible control, compared to using parameters like happiness, while also allowing the use of audio priors. In the following section, I discuss why inviting the users to learn how to effectively communicate with the model is important to achieve this vision.

7.2 Design considerations

As stated at the beginning of this chapter, the aim here is to leverage off-the-shelf text-to-audio models to facilitate the creative expression of novice users, regardless of their musical knowledge. To this end, I embarked upon an examination of potential challenges that these users may encounter during text-to-audio music generation processes to discuss ways to navigate them via interaction design. Guided by the principles of HCI, the think-aloud protocol [332, 5] was utilized by involving three volunteers who self-reported that they possessed no formal musical training beyond compulsory education. Specifically, the

volunteers were provided with access to one of the off-the-shelf text-to-audio models [177] on Google Colab[11] using its official implementation,[12] which enabled them to provide any text prompts and subsequently listen to three music audios generated for each of the text prompts. Here, since the remotely-participated volunteers were Japanese speakers recruited via word-of-mouth communication, they were told that they can use DeepL Translator[13] to translate text prompts into English to obtain better results with the off-the-shelf model that is mainly trained on the dataset with English text labels [90, 149, 73]. They freely used the model for approximately 30 minutes while sharing their screens on a video call and verbalizing their thoughts and feelings, which allowed the identification of the challenges that they encountered and the factors that contributed to these challenges. I then conducted semi-structured interviews and analyzed their responses using open coding [284] to validate the challenges identified and to gain further insight into the reasons behind them. Finally, the following design considerations were yielded by examining how we can construct an interaction design that guides them to overcome the challenges based on Strategy #2 in Section 3.2.

7.2.1 Computational guidance for constructing initial prompts

It was observed that the volunteers frequently encountered difficulty in formulating appropriate text prompts to initiate their use of the model. For example, one volunteer entered the phrase "a song sounds like star wars," resulting in audio containing a battle cry with a space-like sound effect. This can be attributed to the characteristics of the text labels in the dataset used to train the off-the-shelf model [90, 149, 73]. Specifically, the labels of music clips consist primarily of musical descriptions such as genres, instruments, and moods, like: "An orchestra plays a happy melody while the strings and wind instruments are being played [73]." Therefore, providing such a precise description would be essential to ensure that the off-the-shelf model trained on the dataset generates music audio as intended. The volunteer was unable to generate music-like audio until he attempted several prompts and finally entered "solemn music starting with a trumpet fanfare."

In the context of creativity support, two underlying factors could explain the aforementioned observation. First, an inherent gap in artistic vocabulary exists between expert and novice users [290]. Without deep musical knowledge, it can be challenging to conceive a precise description of music audios. Additionally, novice users often have loosely-specified goals when starting a creative endeavor [293, 186], as discussed in Chapter 6. They refine their objectives gradually by exploring the space of possible results through iterative exploration [305, 296]. However, the dependency of off-the-shelf text-to-audio models on precise descriptions of clearly-defined goals makes it difficult for novice users to initiate such exploration. This suggests that supporting them computationally in constructing initial prompts could potentially facilitate the creativity of novice users.

[11] https://colab.research.google.com/
[12] https://github.com/haoheliu/AudioLDM
[13] https://www.deepl.com/translator

7.2.2 Dual-sided iterative exploration of text prompts and audio priors

It was also observed that the volunteers encountered challenges in efficiently exploring the generated results. One volunteer who had prior experience with text-to-image models mentioned the point:

> Unlike text-to-image models, comparing various results at a glance was difficult with the text-to-audio model. So, finding a text prompt reflecting my intention most faithfully became much tough.

In other words, iteratively trying different text prompts would not necessarily assist users in comprehending the space of potential results, although it is vital for novice users to refine their loosely-specified goals [305, 296]. Therefore, users cannot determine which direction would be closest to their goals and what text prompt to try next. Another volunteer mentioned an issue he faced:

> I once found a generation result with a good melody, but I wanted to change its tone. So, I added 'with a flute' to its text prompt and regenerated. However, the melody was then completely changed, which was frustrating.

This implies that we need to let users utilize not only text prompts but also audio priors to constrain the tune of generation results. In sum, supporting the creativity of novice users in music generation processes with off-the-shelf text-to-audio models requires enabling them to efficiently explore variations of both text prompts and audio priors, allowing them to iteratively refine their goals by understanding the space of possible results and acquiring new vocabulary.

7.3 IteraTTA

Based on the above design considerations, this chapter presents IteraTTA, an interaction design dedicated to text-to-audio music generation processes. It was implemented as a Web-based interactive system, allowing novice users to instantly benefit from the off-the-shelf text-to-audio models in their creative processes.

7.3.1 Design

As illustrated in Figure 26, this requires users to first input a theme phrase for music audios to generate. The inputted phrase does not necessarily have to include precise musical descriptions since IteraTTA leverages an off-the-shelf large language model to derive such descriptions suitable for a text-to-audio model using knowledge embedded in such a model [323]. Specifically, the large language model is queried as: "Please give me four variational lists of comma-separated phrases describing what does a music clip of '*[theme phrase]*' sound." It then uses the four responded phrase lists as a variety of the first text prompts to start the music generation processes in parallel. This feature allows novice users to translate loosely-specified goals in their minds into musical descriptions, which can also help them to encounter new vocabulary and envisage variations of text prompts to explore.

IteraTTA then generates three music audios for each of the four prompts. The generated audios are arranged in two dimensions (see Figure 26), which enables novice users to understand how different music audios are generated by different text prompts, and also,

how different music audios are generated by the same text prompts. This is intended to assist users in identifying which text prompts and audio priors are closely aligned with their goals and which direction is worth exploring. If a user identifies a suitable candidate text prompt, they can customize the prompt and generate new music audios with it. Alternatively, if the user discovers a suitable music audio, they can use it as an audio prior to generating new music audios. In essence, the user can explore the subspace of possible results that are proximate to their goals by constraining either text prompts or audio priors, while gradually refining their goals by themselves through understanding the effect of each word in the text prompts on the results.

The presented interaction design has incorporated several features to facilitate the exploration of text prompts and audio priors, as shown in Figure 27. For instance, when a user specifies an audio prior, IteraTTA enables the user to compare generated results with it. It also offers an instant editing feature of text prompts, allowing users to amplify or suppress the sound of a selected instrument. This is achieved by simply adding a phrase of "with strong *[instrument]*" or "with no *[instrument]*" into a text prompt, but it provides an example of how they can modify generation results through prompts.

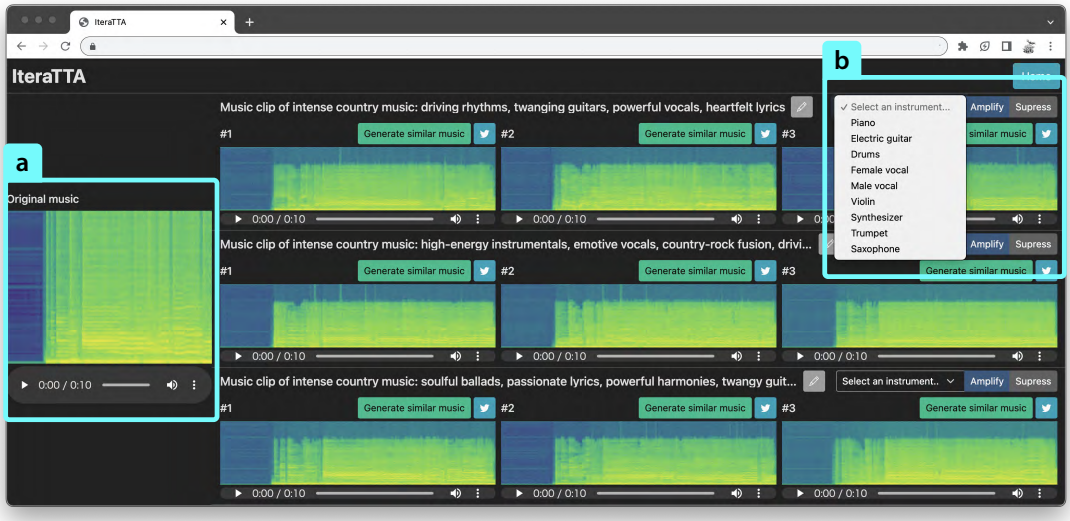


Figure 27: To facilitate the exploration of text prompts and audio priors, IteraTTA allows a) comparison of generation results with an audio prior and b) instant edit of a text prompt.

7.3.2 Implementation

As mentioned above, IteraTTA was implemented as a Web-based system to invite novice users to try music generation with it. The implementation of its back-end server utilized Python with FastAPI and incorporated an API of GPT-3.5[14] to construct initial prompts, while AudioLDM [177] was employed to generate the music audios. The length of music audios to generate was predetermined at 10 s so that a GPU server harnessing an NVIDIA RTX 2080 Ti can afford the generation of 12 audios (3 audios × 4 prompts) simultaneously. On average, the generation process takes approximately 15 s. In addition, DeepL API[15] was used to translate text prompts into English when they were provided in non-English

[14]The `gpt-3.5-turbo` of https://platform.openai.com/docs/models/gpt-3-5 was used.
[15]https://www.deepl.com/docs-api

languages because it was observed that autotranslation led to better results in Section 7.2. The front-end interface of IteraTTA utilized Vue.js, enabling users to download the generated music audios or share them on Twitter.

7.4 Analysis

To investigate the effectiveness of IteraTTA in supporting diverse users in the wild, I deployed it as a publicly available Web-baed interactive system in Japanese. Within two days of release, 8,831 users generated 246,423 music audios. This section discusses the insights extracted from their usage logs and their responses to a form, which was linked from the Web system so that they can share their opinion and feedback voluntarily.

7.4.1 Diversity of theme phrases

I first examined the theme phrases that users inputted to initiate text-to-audio music generation processes and found that they were highly varied. Some users provided music-related phrases, such as "nice city pop" and "cute future bass," while others were more specific, like: "80's hip hop that break dancers would dance to." There were also phrases expressing more abstract ideas, such as "Arabian caves" and "silent dream of a priestess." Figure 28 visualizes the words often used in the translated phrases in the form of a word cloud, showing their diversity.

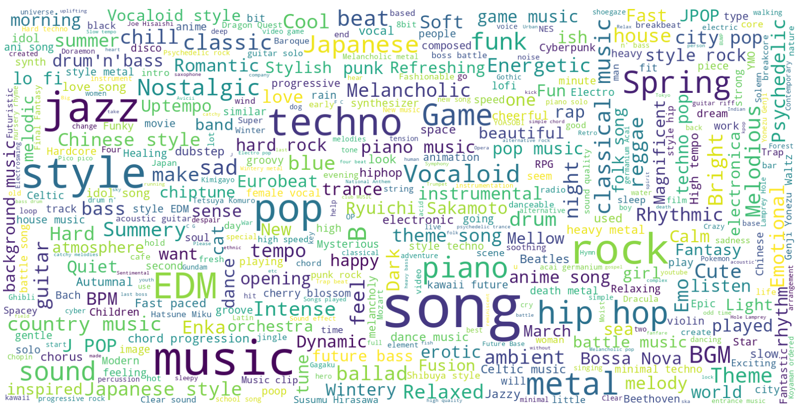


Figure 28: Word cloud of theme phrases the 8,831 users inputted on the Web system.

To explore the role of IteraTTA, the theme phrases inputted by the users and the text prompts derived from them by the large language model were compared to the text labels in the dataset used for training the text-to-audio model. Specifically, 1,000 cases for each of the theme phrases, text prompts, and text labels[16] were sampled to calculate their representation vectors using the same pretrained model of RoBERTa [181] as the text-to-audio model. The distribution of the vectors was then visualized using t-SNE [187], as presented in Figure 29. This indicates that, by leveraging the off-the-shelf large language model to derive text prompts, IteraTTA successfully bridged the gap between the diverse users' theme phrases and the text labels in the training dataset. In fact, it was found that the large language model derived text prompts containing rich musical descriptions even from abstract phrases, such as "otherworldly harmonies, delicate strings, minimalistic percussion, dreamlike vocals" for "silent dream of a priestess." These results suggest the

[16] For the text labels, labels containing "music" were extracted from AudioCaps [149].

effectiveness of guiding the construction of initial prompts to support the creative processes of novice users, as discussed in Section 7.2.1.

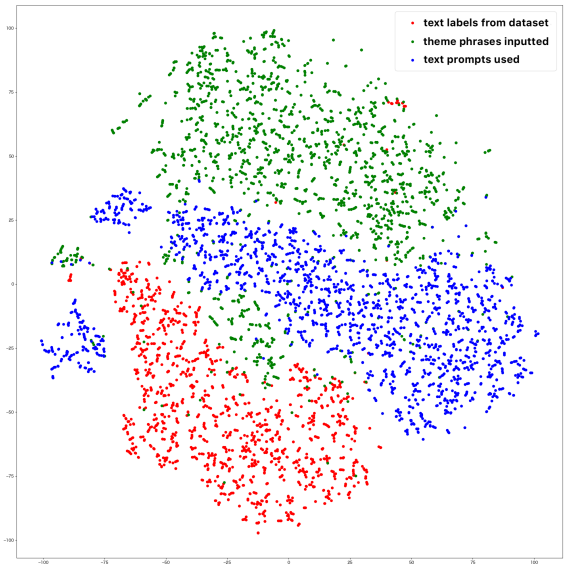


Figure 29: Visualization of the representation vectors of the theme phrases inputted by the users, the text prompts computationally derived from them, and the text labels in the training dataset.

7.4.2 Journey of iterative exploration

I also investigated how the users interacted with the generation results produced by IteraTTA. Specifically, the interaction log of the Web system was analyzed, yielding Figure 30. While some users just tried the exploration feature once, it was found that others made iterative use of the feature, alternating between providing text prompts and audio priors. Interestingly, one user repeated this refinement process 32 times, specifying text prompts 14 times and audio priors 18 times before sharing their final result on Twitter. These points imply that the presented interaction design, which enables dual-sided iterative exploration, helped the users effectively utilize the off-the-shelf text-to-audio model.

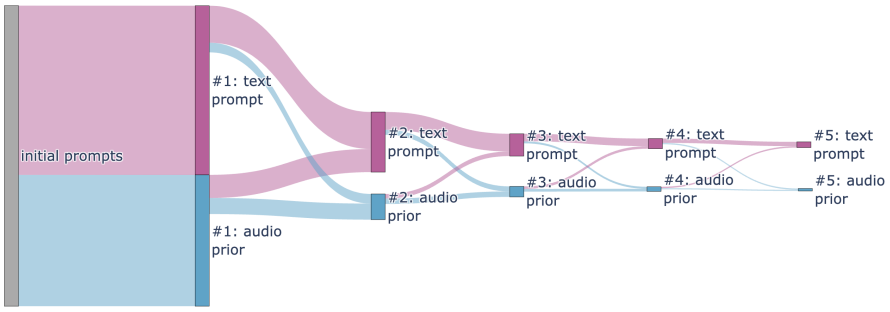


Figure 30: Visualization of how the users utilized the dual-sided exploration of IteraTTA.

7.4.3 Unleashing the creativity of novice users

I lastly analyzed the users' responses to the feedback form, which received 33 responses in total. Overall, most of them expressed their affirmative experiences with the text-to-audio music creation processes, like:

> It was a very interesting trial. I can interact with it throughout the day.
>
> In my personal opinion, it can be used as a source of sampling materials and an idea generator. As a person who usually composes music, I never had any negative feelings about composing from text using this. It is wonderful.

The latter comment suggests that the features of IteraTTA prepared for novice users can also benefit experienced users in different ways. It is also notable that the users left comments implying the importance of the design considerations discussed in Section 7.2, such as how they appreciated the open-ended exploration starting from loosely-specified theme phrases.

> It was fun to encounter songs that fit the theme I provided but I had never heard before.
>
> I really enjoyed the points that I could take advantage of ChatGPT's ability to associate and verbalize even seemingly unconnected ideas, which allowed me to provide crazy theme phrases that would not be understood by a human. I also learned a lot about how to describe songs by looking at the derived text prompts.

These comments specifically highlight the power of interaction designs that can help users learn how to effectively use off-the-shelf ML models.

Interestingly, in the form, some users left a successful prompt that they reached after exploration:

> I would like to report that including a phrase of 'simple progression' or limiting the number of tracks yielded stabilized music audios, like: 'Ideal harmonious song: balanced instrumentation, band sound, simple chord progressions, rhythmic drum patterns, catchy pop melody, up to 12 tracks.'
>
> Adding 'clear sound quality' produces less noisy audios.

It is surprising that, even though no explicit description of the behavior of the off-the-shelf text-to-audio model was provided, the users were able to gain such knowledge by themselves through the iterative exploration with IteraTTA. While such *prompt modifiers* (also known as *quality boosters*) [223] that influence results in a specific way have been discovered for text-to-image models in a community-driven manner [223, 224], the above comments would be the first examples for text-to-audio models, to the best of our knowledge. We can conclude that this is a manifestation of users' creativity in text-to-audio music generation processes and would be hard to derive without IteraTTA, specifically, its design to foster users' learning.

7.5 Summary

This chapter introduced IteraTTA, an interaction design specifically dedicated to supporting novice users, who do not necessarily have rich musical vocabulary, in their text-to-audio music generation processes. This is guided by the strategy of fostering users' understanding of how to effectively use off-the-shelf text-to-audio models and built on two main features a) computational guidance for constructing initial prompts and b) dual-sided iterative exploration of text prompts and audio priors. The former can help novice users translate

their loosely-specified goals into text prompts, which serve as starting points for exploration while providing an opportunity to encounter new vocabulary. The latter is important for enabling them to comprehend the space of possible results and the effect of each word in text prompts on the results, helping gradually refine their goals. How diverse users can utilize IteraTTA in their creative processes was confirmed by deploying it as a publicly available Web-based interactive system and analyzing users' behaviors, which highlighted the importance of these design considerations in supporting the users' creativity. In combination with the findings from Chapters 4 to 6, this result suggests that Strategy #2 outlined in Section 3.2 is capable of generating effective interaction designs using off-the-shelf ML models to address diverse user needs, thereby confirming RQ1 and RQ2.

8 BeParrot: An interaction design for training users to transcribe unclear speech with off-the-shelf speech recognition models

This chapter presents the last example in this dissertation, which addresses the challenge in speech transcription with off-the-shelf ML models. Needless to say, the importance of speech transcription is widely acknowledged because the acquired text can be used in diverse situations, *e.g.*, searching audio content [220], increasing the accessibility of videos [117], analyzing language production [200], and developing various speech processing methods such as speech recognition [209] and voice conversion [276, 8]. However, manual transcription is time-consuming and tedious, and thus, previous studies have attempted to address this issue by utilizing ML technologies. Specifically, a post-correction approach has been applied to handle this issue, in which an automated speech recognition (ASR) model is first applied to the speech to transcribe, and then a human corrects errors in the recognition results while listening to the audio. It has been confirmed that the improvement of ASR models allows people to correct fewer errors in the recognition results, thereby reducing the total time for transcription [179, 185].

In other words, the efficiency of the post-correction approach is highly dependent on the accuracy of ASR models. In particular, this approach is known to be less effective when the performance of the ASR models is low, *e.g.*, when the audio to transcribe is unclear due to its recording condition. Gaur *et al.* [88] and Sperber *et al.* [280] experimentally demonstrated that the word error rate (WER)[17] is desirable to be less than 30% for the post-correction approach to be effective. Therefore, previous studies have focused on applying this approach to the transcription of clear speech, *e.g.*, audio from TED Talks [88, 280], news [316], and lectures [201]. Given that transcribing unclear speech (*e.g.*, speech with a lot of noise or reverberation) is still difficult for most ASR models [281, 302, 105], it can be expected that this post-correction approach does not effectively help users in such cases. While it is feasible to customize ASR models for improved accuracy, this involves not only the preparation of a dataset that has been recorded under the same noise and reverberation conditions and transcribed beforehand [179, 280] but also the challenges discussed in Section 2.3.

To overcome this challenge, this chapter presents *BeParrot*, an interaction design that enables efficiently transcribing unclear speech with off-the-shelf ASR models. Informed by Strategy #2 presented in Section 3.2, it is designed to foster the skills of users in *respeaking*. Respeaking is the method that has primarily been used to create subtitles for television programs in real time [135, 194, 252]. Rather than directly inputting the audio of the programs into an ASR model, respeaking involves a human *respeaker* who repeats the speech of the programs as if speech shadowing [195]. The respeaker's utterance is inputted into an ASR model that outputs the corresponding transcription in real time, which is used as subtitles after the respeaker or another person manually corrects errors as necessary. This method is effective when the ASR model can recognize the utterance of the respeaker more accurately than the original speech in television programs, thereby reducing the number

[17] WER denotes the percentage of words that are not correctly recognized. The lower value of WER indicates that the recognition results contain fewer errors and are more accurate.

of required manual error corrections. Analogously, we can expect that respeaking can be used to improve the efficiency of transcribing unclear speech. This is because, if people can repeat the content of speech with noise or reverberation in a quiet environment with a clear voice, their utterance will be recognized accurately using an off-the-shelf ASR model. Then, this reduces the number of errors in the recognition results to be post-corrected, compared to directly inputting the unclear speech into the ASR model.

On the other hand, it is known that the effectiveness of respeaking is highly dependent on the proficiency of the respeaker [252, 253]. In particular, respeakers incur high cognitive demand, as they are required not only to listen to what is being said in order to repeat it without delay but also to simultaneously memorize the speech content for post-correction of errors in the recognition result. In addition, respeakers should be able to utter clearly without stuttering or stammering so that utterances are transcribed accurately by the ASR models. These points necessitate BeParrot to be carefully designed to help users get used to respeaking, as suggested by Strategy #2 of Section 3.2. To this aim, BeParrot utilizes the history of how a user has interacted with its interface (*e.g.*, retrying utterance and editing recognition results), which would reflect the user's ability of respeaking, to provide two key features in BeParrot: *parameter adjustment* and *pronunciation feedback*. The parameter adjustment feature automatically updates two parameters, the playback speed and the length of each speech segment, because these parameters determine the difficulty of respeaking according to previous studies [252, 253, 251]. For example, when the user has retried an utterance of the same segment, the user is likely to have trouble in respeaking the segment, and then, the playback speed can then be automatically decreased. The user of BeParrot can also manually adjust these parameters when they become accustomed to respeaking and want to increase the playback speed or segment length. Furthermore, the pronunciation feedback feature presents phonemes that are difficult to be recognized by the off-the-shelf ASR model when pronounced by the user, which is derived from the history of their manual error corrections. Then, the user can leverage the feedback to improve their pronunciations, especially of those the ASR model would not recognize correctly.

In this chapter, the effectiveness of BeParrot on different types of speech data was evaluated by involving 60 crowd workers. The results successfully demonstrated that the workers could transcribe the speech more efficiently with BeParrot than with a conventional post-correction approach using the same off-the-shelf ASR model. Particularly, the effectiveness of BeParrot was confirmed when the speech is unclear and difficult to recognize using ASR models, *e.g.*, speech with high levels of noise or reverberation [281, 302, 105]. In addition, comments from the workers qualitatively supported the design of BeParrot, expressing their affirmative reception of both the parameter adjustment and pronunciation feedback features. At the same time, this confirms the effectiveness of the strategies outlined in Section 3.2 in supporting speech transcription with off-the-shelf ML models.

8.1 Related work

BeParrot is designed to address the user need for efficiently transcribing speech regardless of the recording condition. To situate it, this section first reviews previously proposed interactive systems for supporting speech transcription, most of which adopt the post-correction of the recognition results of off-the-shelf ASR models. Then, the concept of

respeaking is described, especially how it has been used and the difficulties novice users encounter when performing respeaking.

8.1.1 Interactive systems for supporting transcription

To moderate the importance of speech transcription and its time cost, there are several studies that propose supporting interactive system for transcription tasks, as mentioned at the beginning of this chapter. For example, as one of the initial works, Barras *et al.* [18] proposed an integrated text editor that visualizes a speech wave. Given the development of ASR technologies, the post-correction approach has become widely used to assist speech transcription [179, 185, 184, 201, 280, 316]. For example, Liu and Soong [179] developed a handwriting user interface that allows users to correct errors in a convenient manner. Luz *et al.* [184] developed a 3D game-based interactive system to support a collaborative correction process by motivating users. These approaches assume that the speech to transcribe can be recognized by ASR models with certain accuracy. For example, the WER of the pre-correction transcription was reported to be 21.5% in the study conducted by Luz *et al.* [185] and 28.8% in the study conducted by Miro *et al.* [201]. This setting is in agreement with the findings of Gaur *et al.* [88] and Sperber *et al.* [280], who concluded that the WER is desirable to be less than 30% for the post-correction approach to function.

However, depending on the nature of the audio source, this is not always feasible, even with concurrent off-the-shelf ASR models. Specifically, the performance of the ASR models is degraded when the speech is unclear, *e.g.*, when the speech contains high levels of noise or reverberation. For example, Tsunoo *et al.* [303] reported a WER of 48.6% with their ASR model on a dataset of unclear speech [322] even though they leveraged various data augmentation techniques involving reverberation simulation and adversarial training. Yet, previous studies for supporting speech transcription often evaluated their approach using clear speech, *e.g.*, audio from TED Talks [88, 280], as mentioned at the beginning of this chapter. By preparing a dataset collected under the same recording condition, it is possible to customize ASR models to achieve high accuracy against unclear speech [179, 280], but it imposes various difficulties mentioned in Section 2.3 and limits their applicability to a wide variety of speech data. These limitations demanded a better interaction design that can work without dependence on the audio source.

8.1.2 Respeaking

As mentioned at the beginning of this chapter, respeaking has been practically used by many broadcasting stations, *e.g.*, BBC [194] and NHK[18] [135, 116], to create subtitles for news programs or sportscasts in real time. To minimize the delay when creating subtitles, respeakers must repeat the speech clearly without stammering such that the repeated speech can be transcribed accurately by an ASR model. Therefore, broadcasting stations assume professional respeakers who have completed specialized training [252, 253]. For example, Prazak *et al.* [236] described how Czech Television leverages respeaking, stating that respeakers are required to undergo 75 hours of in-house training. In addition, Waes *et al.* [312] found that respeakers are required to employ a variety of strategies to limit information loss as much as possible.

[18]NHK is a Japanese government-owned public broadcasting station.

Given that respeaking demands high proficiency, previous studies under-explored the possibilities of novice users transcribing speech via respeaking. In fact, the difficulties associated with mastering respeaking were discussed by Ghyselen *et al.* [92] as a reason why they did not adopt this approach for transcribing a dialect corpus in their study. In addition, Sperber *et al.* [279] empirically demonstrated that the effectiveness of respeaking is strongly dependent on how accurately the respeaker's utterances are recognized in a study where two speakers tried transcribing TED Talks via respeaking. Vashistha *et al.* [308, 307, 306] employed a workaround in their studies by the introduction of segmenting and majority-voting. Here, they deployed crowdsourcing transcription tasks in which speech data were divided into small segments (less than 6 s), and the speech content of each segment was uttered by five crowd workers via respeaking. The final transcription was obtained via a majority vote among the recognition results of the repeated speech. Given that, we can expect that respeaking can be a powerful tool for speech transcription if it is made easier for novice users. In particular, it would be especially efficient for transcribing unclear speech, which is typically challenging using the conventional post-correction approach (see Section 8.1.1), because respeakers will be able to utter clear speech that is recognizable by off-the-shelf ASR models without depending on the quality of the original audio. Therefore, informed by the strategies in Section 3.2, this chapter introduces a new interaction design that helps novice users utilize and get accustomed to respeaking.

8.2 BeParrot

This section explains the design of BeParrot, starting by explaining its two key features that help novice users perform respeaking based on the findings of previous studies on the training programs of respeaking. Subsequently, the implementation details of BeParrot are presented.

8.2.1 Design

In speech transcription tasks, an audio clip to transcribe is typically divided into short segments in advance, and each segment is transcribed sequentially by users. Similarly, BeParrot assumes that each segment is transcribed via respeaking along with manual correction. However, as mentioned in Section 8.1.2, respeaking is a highly demanding task that is difficult for novice users. Thus, prior to developing BeParrot, I referred to studies on training programs for professional respeakers [252, 253, 251]. Based on the findings of these studies, two key features were constructed so that they would help novice users utilize respeaking without requiring extensive training.

The first feature is parameter adjustment, which allows users to control both the playback speed of the speech and the length of each speech segment dynamically during the task. Relative to playback speed, Fresco [251] reported that, in some respeaker training programs, trainees attempt to identify an optimal playback speed through multiple steps. They stated that the optimal speed varies significantly for different people, and the speed affects respeaking performance. These observations informed the adjusting feature of the playback speed. In addition, the length of each speech segment was made adjustable to allow users to control their cognitive load during the task. In fact, Fresco [251] reported that trainees identified "multitasking" as the most difficult aspect of respeaking. Specifically,

respeakers must clearly repeat what is being said while listening to and remembering the speech, as discussed at the beginning of this chapter. Thus, it can be anticipated that allowing users to adjust the segment length based on their ability would be helpful.

In addition to making these two parameters adjustable by users, BeParrot is designed to automatically adjust the parameters based on users' interaction history. Specifically, if a user is experiencing difficulty in respeaking (*e.g.*, retrying a specific speech segment over and over), BeParrot automatically reduces the playback speed and segment length. This is because the effectiveness of such automated adjustments has been confirmed previously in the development of a tool to support language learning via speech shadowing [369].

The second feature is pronunciation feedback, where users are prompted to be careful about pronouncing specific phonemes. Given the nature of respeaking, utterances should be recognized accurately by the off-the-shelf ASR model, and respeakers must utter clearly without stuttering or stammering [252]. Thus, this feature is prepared to increase awareness of certain phonemes in utterances that are likely to be missed or misrecognized by the ASR model. This feedback feature can be achieved by analyzing how users manually correct the recognition results, which is described in detail in the next section.

8.2.2 Implementation

To realize these two key features, BeParrot was implemented in the form of a Web-based interactive system (Figure 31). I implemented its interface in Japanese because a user study is conducted on a Japanese crowdsourcing platform,[19] as described later in Section 8.3.

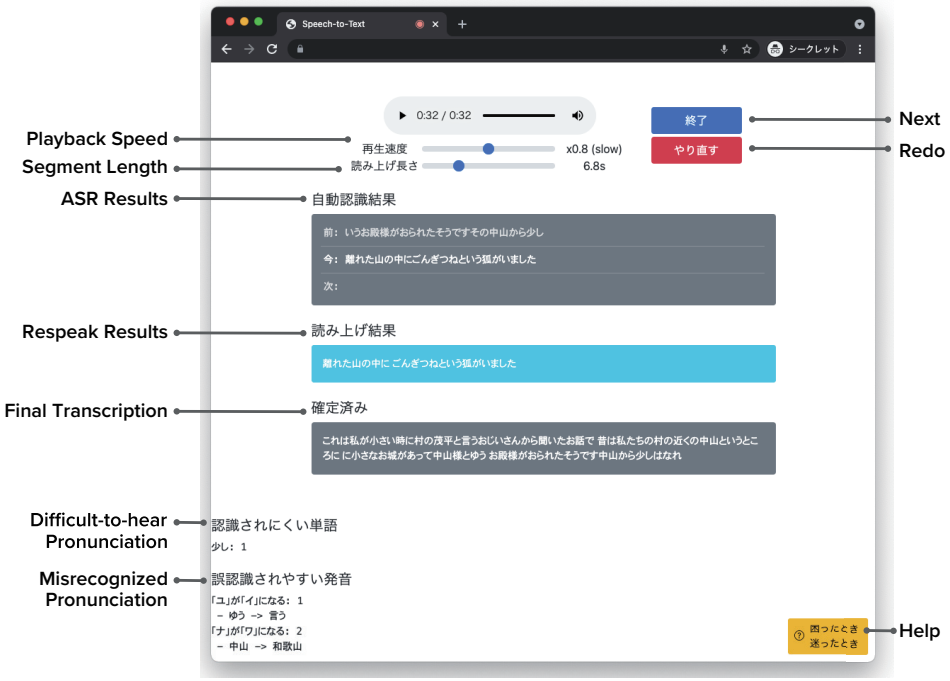


Figure 31: BeParrot interface. The interface was implemented in Japanese, and each text outside the window indicates corresponding meanings in English.

As mentioned in Section 8.2.1, the entire audio clip to be transcribed is first divided into short pieces using a voice activity detection technique [277], specifically webrtcvad.[20]

[19] https://www.lancers.jp
[20] https://github.com/wiseman/py-webrtcvad

The split speech pieces, which are approximately from 0.1 to 3 s, are then concatenated into a segment until its length exceeds the parameter specifying the length of each speech segment using a greedy algorithm. Once the parameter is adjusted either manually or automatically, the concatenation process is re-executed. In BeParrot, a user is supposed to transcribe each of the concatenated segments in order via respeaking. Here, the user's utterance of a segment is recognized sequentially using a streaming ASR model, and the recognition result appears in "Respeak Results" in real time. Note that the user can edit errors in the text manually using a keyboard if necessary. Once the transcription for the segment is finalized by clicking "Next", it is concatenated to the entire transcription shown in "Final Transcription". Otherwise, if the user wants to utter the same segment again, they can do so by clicking "Redo". In addition, in a similar manner to the post-correction approach mentioned in Section 8.1.1, the audio source to transcribe is recognized using an ASR model in advance, and the result is presented in "ASR Results" for reference. Here, the user can observe three transcriptions in a row that correspond to the speech of the last, current, and next segment, respectively. In addition, there is a "Help" button in the bottom-right area that the user can press at any time to view information about the usage of BeParrot.

To achieve the parameter adjustment feature as described in Section 8.2.1, two sliders were implemented, *i.e.*, the "Playback Speed" and "Segment Length" sliders, to allow users to adjust the playback speed and segment length, respectively. Note that these parameters are also adjusted automatically based on the interaction records of the individual user. Specifically, if the user retries uttering the same segment multiple times, BeParrot identifies the user as having trouble uttering it without stuttering or stammering and slows the playback speed by $\times 0.85$. If the user stops playback in the middle of the segment, the user may encounter "multitasking" difficulties with the long segment. Thus, BeParrot automatically reduces the length of consecutive segments to the length they stopped at with the decay parameter of 0.5.

The bottom-left area of BeParrot is used for the pronunciation feedback feature introduced in Section 8.2.1. In "Difficult-to-hear Pronunciation", words that are not recognized by the streaming ASR model and added manually by the user using the keyboard are listed in the order of the number of additions. In addition, "Misrecognized Pronunciation" presents a list of moras that are often misrecognized by the ASR model based on how the user corrected the recognition results of their utterances. This list is obtained by calculating the optimal edit operations between the mora sequences of the original and corrected word using the Wagner–Fischer algorithm [313]. The calculated edit operations over speech segments allow BeParrot to identify how many times each mora was corrected and to rank those that are frequently misrecognized.

Here, two ASR models are employed in BeParrot: one processes the audio clip in advance, and the other processes the user's utterances sequentially in real time. For the former model, this implementation used an off-the-shelf model of Conformer [101], which is one of the state-of-the-art ASR models based on both Transformer and CNN architectures. The Conformer model was trained on a Japanese corpus using ESPNet [321], which is an open-source end-to-end speech processing toolkit. For the latter ASR model, it used

Google Speech-to-Text API,[21] which allows us to transcribe streaming audio in real time. The latency of respeaking (*i.e.*, from the time a user utters to the time the corresponding recognition result appears on "Respeak Results") was approximately 800 ms. Note that, when using BeParrot, the user is required to wear headphones to avoid the original speech from being mixed with the user's utterances.

8.3 User Study

To confirm the effectiveness of BeParrot, a user study was conducted involving crowd workers. The workers were asked to transcribe different types of speech data using BeParrot or the conventional post-correction approach. Then, the time they spent in transcribing the same speech and the accuracy of the obtained transcriptions were compared between the two approaches, *i.e.*, respeaking and post-correction.

8.3.1 Materials

This user study used three types of speech data: *clear*, *radio*, and *historical* speech. Each type comprised two audio clips of three minutes, bringing six clips in total. Here, since the crowd workers were recruited in Japan (see Section 8.3.2), the clips were prepared in Japanese. The clear clips were taken from TED Talks recorded in a quiet environment. The radio clips were prepared to evaluate the effectiveness of BeParrot on noisy speech because they were taken from a radio broadcast in which two professional speakers debated while background music was played. The historical clips were more challenging because they were taken from two lectures from the 1970s. These clips contain high levels of reverberation as well as noise because the lectures were recorded using a single microphone located in a large auditorium. For each clip, its transcription text was also prepared as the ground truth from the corresponding source (*e.g.*, the TED Website for the clear clips) to evaluate the accuracy of transcriptions obtained in this study. Using the ground truth, it was confirmed in advance that the clips were sufficiently unclear to transcribe using the off-the-shelf ASR model such that they resulted in a character error rate (CER) of more than 30%,[22] as shown in Table 23.

8.3.2 Design

A between-participant design was employed across the *baseline* and *proposed* conditions. Here, I did not include a condition where a participant performs transcription tasks via respeaking without the support of BeParrot, so to speak, *vanilla* condition. This is because, in the initial exploration, it was found that novice users could not conduct respeaking in such a *vanilla* condition, which is attributed to the difficulties associated with conducting respeaking (see Section 8.1.2).

For the baseline condition, a Web-based system was prepared to allow the crowd workers to transcribe speech using the post-correction approach. Similar to the system used by Gaur *et al.* [88], the Web-based system comprised an audio player and presentation of

[21] https://cloud.google.com/speech-to-text/docs/streaming-recognize

[22] As mentioned in Section 8.3.3, the value of WER generally correlates with the value of CER, and thus, the higher value of CER indicates that the recognition results contain more errors. In addition, the value of WER is usually higher than the value of CER.

the recognition results obtained by the ASR model. Its interface also replicated keyboard shortcuts that Gaur *et al.* [88] implemented to allow the workers to play, pause, or rewind 5 s. After being presented the instructions on how to use this system, the workers assigned to the baseline condition were asked to transcribe one of the six prepared clips. For the crowd workers assigned to the proposed condition, BeParrot was provided to transcribe the speech. Since it was expected that workers are unaccustomed to respeaking, they were asked to first practice the use of BeParrot by transcribing a clear speech of 30 s[23] after watching an instructional video. This process took approximately 3.5 minutes on average. Then, the workers were asked to transcribe one of the prepared clips in the same manner as those in the baseline condition. Here, the implementation of BeParrot was customized to record workers' interactions, *e.g.*, retrying an utterance, changing playback speed, and correcting errors using a keyboard.

In addition, the workers were asked to complete a questionnaire after they finished the transcription task. The questionnaire included the items from NASA-TLX [107, 37] to compare the workers' cognitive loads between the two conditions. In addition, for the workers assigned to the proposed condition, there were questions to collect their opinions regarding the usage of BeParrot, *e.g.*, "*how useful was the parameter adjustment feature?*", "*how useful was the pronunciation feedback feature?*", and "*please write down anything else you noticed about the experience of using this interface*". Their responses to these questions were later analyzed using open coding [284] to enumerate major topics.

For each of the two conditions, 30 crowd workers were recruited and randomly assigned to one of the six clips, resulting in five workers for each clip. The recruitment process was performed on a Japanese crowdsourcing platform, and they were paid approximately $10 for their participation. For the proposed condition, the workers were required to use headphones or earphones (not speakers) so that they could perform respeaking without trouble, as mentioned in Section 8.2.2. In addition, to exclude data of workers who have an experience of respeaking in the past, two of the authors independently examined the questionnaire responses from the workers of the proposed condition, but no cases were confirmed.

8.3.3 Measure

To evaluate the effectiveness of BeParrot, two measures were prepared: time and CER. Specifically, the time each worker spent transcribing clips of the same length was measured and compared across the two conditions to verify whether BeParrot contributed to the efficiency of the transcription task. The CER of the obtained transcriptions was also compared to confirm the effect of BeParrot on transcription accuracy. Note that CER is widely used to evaluate the quality of transcription in languages without space delimiters [162], including Japanese [120], because the WER value calculated for such languages depends on the quality of morphological analysis. Still, the value of CER generally correlates with the value of WER while it is expected to be lower than WER, as in Petridis *et al.* [229].

Table 22: Time crowd workers spent transcribing. The proposed condition significantly shortened the time spent for the radio and historical clips.

Speech type	Baseline (s)	Proposed (s)	Reduction (%)
Clear	1632.7 (±224.6)	1505.8 (±426.8)	7.8
Radio	2758.1 (±273.6)	1823.2 (±253.4)	**33.9**
Historical	2004.6 (±205.0)	1408.4 (±140.4)	**29.7**

Table 23: CER of the obtained transcription. No significant difference was found between the baseline and proposed condition.

Speech type	ASR	Baseline	Proposed
Clear	6.16	3.73 (±0.57)	5.75 (±1.02)
Radio	30.72	19.05 (±1.56)	24.81 (±2.69)
Historical	48.18	29.25 (±5.16)	29.10 (±2.17)

8.3.4 Results

The results are shown in Table 22 and Table 23. According to the t-test, it was found that the workers assigned to the proposed condition spent significantly less time transcribing the radio ($p = 0.037, d = 1.01$) and historical ($p = 0.026, d = 1.08$) clips. On the other hand, no significant difference was found in CER for all clips. These points imply that BeParrot contributed to the reduction of the time required to transcribe unclear speech (32.1 % in total for the radio and historical clips[24]), compared to the conventional post-correction approach, without significantly deteriorating transcription accuracy. The fact that the time required to transcribe clear speech was not significantly different reconfirms the findings of previous studies suggesting the effectiveness of the post-correction approach for clear speech, as mentioned in Section 8.1.1.

In addition, it is notable that the baseline condition showed a significant increase in the time required to transcribe unclear speech ($p = 0.004$) according to the one-way ANOVA. Here, the recognition results for the radio and historical clips obtained by the ASR model that were presented to the workers exhibited a CER value that was greater than 30 %. Thus, the time increase can be said to be analogous to previous studies [88, 280], which demonstrated that recognition results in the post-correction approach should be sufficiently accurate (*e.g.*, less than 30 % WER). On the other hand, the proposed condition did not exhibit a significant difference in the time spent across the three types of speech. This highlights the effectiveness of BeParrot, as it did not exhibit a time increase against such unclear speech.

Also, the workers' responses for the NASA-TLX items were compared between the two conditions in the same manner as Zhang *et al.* [369]. The results are shown in Figure 32, where the higher score indicates a higher stress level for each of the six items. In total, no significant difference was found in the obtained scores, which implies that BeParrot allowed the workers to complete the transcription task without requiring extra stress compared to the conventional post-correction approach. However, the scores indicate that the amount of

[23] Note that the clip of clear speech used for the practice was a recitation of a famous children's story and independent of the six clips prepared in Section 8.3.1.

[24] $100 - \frac{1823.2\,(\text{radio, proposed}) + 1408.4\,(\text{historical, proposed})}{2758.1\,(\text{radio, baseline}) + 2004.6\,(\text{historical, baseline})} = 32.14_{(\%)}$

physical effort required for the transcription task was reduced with the proposed condition. This suggests that respeaking is less physically demanding compared to correcting erroneous recognition results using a keyboard.

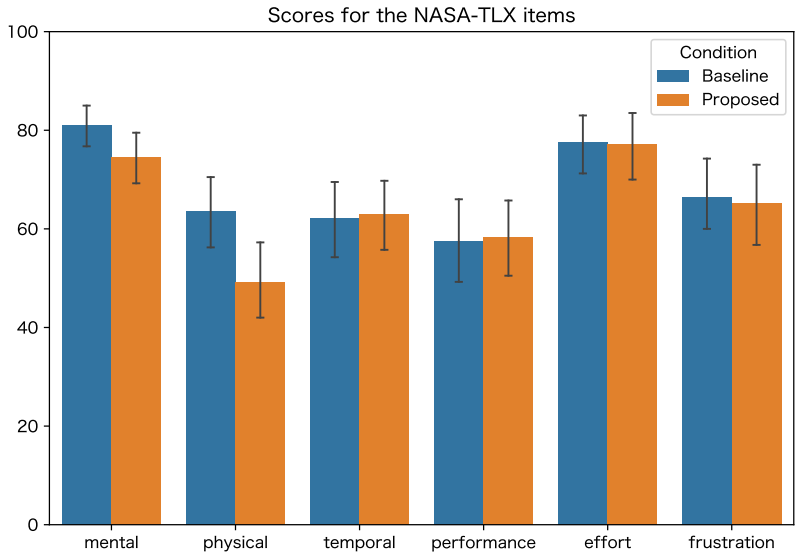


Figure 32: Comparison of worker cognitive load based on NASA-TLX (error bars indicate 95% confidence intervals).

Furthermore, the interaction logs of the workers suggested the effectiveness of the parameter adjustment feature. For example, Figure 33 shows the transition of the playback speed that the workers assigned to the same clip experienced as a result of the manual and automated adjustment. As can be seen, the playback speed was gradually reduced automatically until it reached a specific value that seemed to be an optimal value for each worker. In other words, the playback speed BeParrot stopped slowing down (*i.e.*, the speed that allowed each worker to perform respeaking without retrying) differed for each worker. This confirms the validity of its design, which was based on the report by Fresco [251], as described in Section 8.2.1. In addition, from the fact that the worker denoted by the purple solid line increased the playback speed at the last part of the task, we can infer that the worker got familiar with respeaking and did so. While further discussion is presented in Section 8.3.5, these results imply the effectiveness of the design of BeParrot, especially regarding its power of fostering users' skill of respeaking.

8.3.5 User comments

As explained in Section 8.3.2, the crowd workers assigned to the proposed condition were asked to complete a questionnaire. According to their responses, they overall considered BeParrot to be useful and expressed their willingness to continue using it.

> This was my first time using this kind of transcription tool, but once I got used to it, I was able to transcribe a little faster, which I found very useful.
>
> I felt that using the voice input function would make the transcription process smoother and less stressful.
>
> It was very easy because what I had to do was limited to minor adjustments as long as I focused on listening.

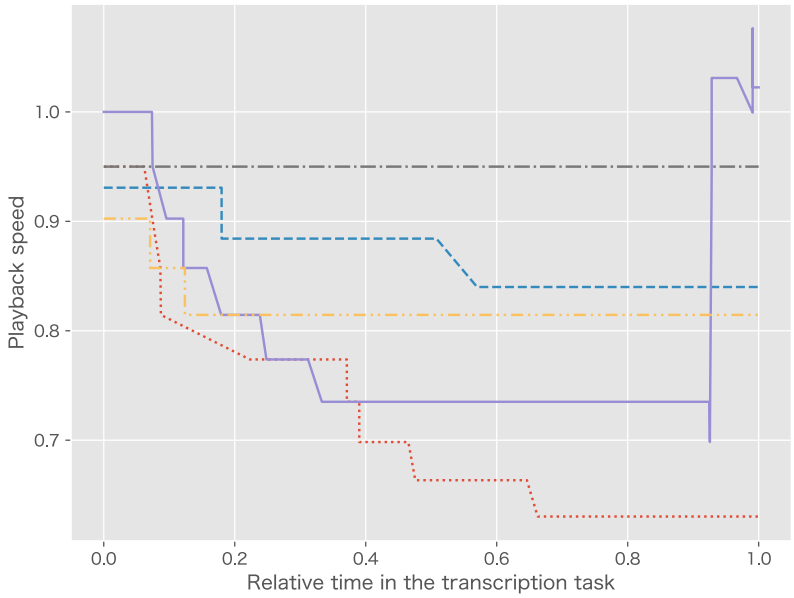


Figure 33: Transition of playback speed in the relative time of the transcription task (each line represents the transition of different workers assigned to the same clip).

In addition, the key features of BeParrot described in Section 8.2.1 received positive comments. In particular, the workers frequently commented on the parameter adjustment feature. For the playback speed, some workers indicated the usefulness of the control slider, and others mentioned that the automated speed adjustment was effective.

> Every time I retried the transcription, it automatically slowed down the playback speed, which saved me from having to do it myself and made the task easier.
>
> The slowing down of the playback speed made it easier to listen and was very useful for correcting mistakes.
>
> It played faster the first time and slower since the second time, which was helpful in allowing me to transcribe more efficiently and accurately.

The responses indicated that, as well as the adjustment of the playback speed, the adjustment of the segment length eased the complexity of respeaking.

> When the segment was long, I had my hands full memorizing the content, and it was difficult to read it aloud simultaneously. Thus, I actually utilized the adjustment feature. When dealing with complicated contents or unfamiliar fields, I think that the length adjustment function is indispensable to make sure that I can understand each phrase. Then, the relistening to long segments can be avoided, and the time required in total can be shortened.
>
> I felt it very difficult to transcribe a long segment. The system then automatically adjusted the length of the segment to be shorter, and after that, the transcription process became much easier.

These responses confirm that the parameter adjustment feature helped novice users overcome the difficulty of respeaking, supporting the design of BeParrot described in Section 8.2.1. Note that this feature is newly introduced for respeaking with BeParrot, informed by the strategies of Section 3.2, to help users get accustomed to it. Previous applications

of respeaking for the subtitle creation for television programs, which were mentioned in Section 8.1.2, did not allow such a feature since changing the playback speed or retrying the same segment makes it impossible to create subtitles in real time.

For the feedback feature, some workers said that it helped them become more aware of their pronunciation habits.

> The presentation of the number of incorrect pronunciations was very helpful. I realized that I have a bad tongue and I needed to be more careful in my daily life.

> It was useful because it made me aware of my pronunciation, especially about the "la" and "na" columns *[moras starting with the consonants of "l" and "n"]* that were displayed as often mispronounced.

These responses further corroborate the importance of the interaction design that fosters human learning and its potential to facilitate efficient collaboration with off-the-shelf models. However, some responses suggested that monitoring the feedback while respeaking was difficult.

> I was engrossed in uttering what I heard and did not look at the feedback at all.

> My mind was so occupied with the transcription task that I did not think to refer to it much.

One worker mentioned that they had trouble improving their pronunciation from the display.

> The presentation of pronunciations that tend to be misrecognized let me carefully utter them. Still, they were not recognized accurately.

These responses suggest that there is room for improvement regarding the presentation of the feedback. For example, showing the feedback after a user completes a transcription task is a possible option to ease the difficulty of referring to the feedback during respeaking. This would give them time to reflect on the feedback and consider how to improve their pronunciation in future tasks. For example, computationally providing auditory feedback might also be helpful for them to improve their pronunciation [205].

8.4 Limitations

The results presented in Sections 8.3.4 and 8.3.5 demonstrate the effectiveness of BeParrot as a tool for transcribing unclear speech using off-the-shelf ASR models. The results also indicate that its design (*i.e.*, the parameter adjustment and pronunciation feedback) helped novice users perform respeaking. On the other hand, we can consider several limitations. First, as the key component of BeParrot, there was an assumption that the accuracy of the off-the-shelf ASR model that recognizes respeakers' utterances should be high. This limits the applicability of the presented interaction design, as it can be difficult to use BeParrot for a specific type of speech, for which accurate ASR models have not been developed (*e.g.*, speech in languages with low resources). Consequently, evaluations in various languages and cultural backgrounds would be demanded to increase the generality of the results [175]. In addition, despite the efforts to use different types of audio clips, as mentioned in Section 8.3.1, it is desirable to evaluate the effectiveness of BeParrot when it is used to transcribe more diverse types of speech. For example, given the nature of

respeaking, the difficulty of using BeParrot may increase when the speech to transcribe contains spontaneous conversation by multiple people with many overlaps. The effectiveness of BeParrot in transcribing longer speech (*e.g.*, over an hour) is also worth examining, as it will further clarify its longitudinal learning effect, which includes both that a user becomes increasingly familiar with the interactive system and that the user improves their pronunciation based on the feedback feature. Longer transcription tasks would also reveal the effect on users' cognitive load in comparison with conventional approaches.

8.5 Summary

This chapter presented BeParrot, an interaction design constructed for efficiently transcribing unclear speech via respeaking with the help of off-the-shelf ML models. Guided by Strategy #2 of Section 3.2, it features parameter adjustment and pronunciation feedback to enable novice users to conduct and get accustomed to respeaking without extensive training. A user study involving 60 crowd workers examined how BeParrot helps users transcribe different types of unclear speech, such as speech with high levels of noise or reverberation. The results demonstrate that, compared with a conventional approach, BeParrot makes transcription tasks of unclear speech 32.2 % faster without losing the accuracy of transcriptions. From this, we can infer that transcribing unclear speech should still be hard for either a human or an off-the-shelf model alone. At the same time, BeParrot would be an illustrative example showing that, even in such situations, humans and off-the-shelf models can work effectively by preparing an interaction design that guides humans to adapt their skills. This point affirms RQ1 and RQ2 of Section 3.4, given that the strategies generated the effective interaction design for speech transcription.

9 Discussions

So far, I have introduced five different interaction designs, all of which were designed based on the strategies presented in Section 3.2 for the purpose of reconciling off-the-shelf ML models with open-ended user needs. Based on the observations and findings from each example, this chapter first provides a reflection on the research questions in Section 3.4. This outlines the effectiveness of the proposed approach that focuses on the plasticity of user needs and behaviors and conceptualizes the relationship between off-the-shelf models and end-users as a bivariate problem. Also, this chapter discusses the limitations of the proposed approach, such as ethical considerations, and its future directions to ensure that it serves as a *concept* [282, 119] capable of generating new interaction designs.

9.1 Answers to the research questions

Table 24 presents the summary of the interaction designs presented, outlining how each design addresses specific user needs using various off-the-shelf ML models. As discussed in the respective chapters, these designs are based on the strategies introduced in Section 3.2, and their effectiveness has been confirmed in different forms. It is notable that no training of ML models has been involved to realize these diverse interaction designs. Returning to the research questions posed in Section 3.4, it now seems appropriate to reflect on how these designs align with and answer that initial inquiry.

> **RQ1**: Can the above strategies generate diverse interaction designs against open-ended user needs using off-the-shelf ML models?

To this question, we can conclude that the five examples presented an affirmative response. Notably, by employing two different strategies, it was demonstrated that, even when an appropriate human cognitive process to be utilized is absent, fostering human learning can be an alternative way to generate interaction designs against different tasks across multiple domains. While it is infeasible to deductively demonstrate that these strategies are truly applicable to all open-ended user needs, I believe they are sufficiently robust to serve as a generative concept to be referenced by developers and practitioners.

The five examples also affirmatively answer RQ2.

> **RQ2**: Can the interaction designs generated from the strategies effectively address user needs by using off-the-shelf ML models?

Here, each example employed a different evaluation scheme, ranging from user studies and crowdsourcing-based evaluations to in-the-wild studies. Despite their variety, all these evaluations aimed to ascertain the effectiveness of each interaction design, confirming its capability of addressing user needs as well as some limitations. This suggests that the two strategies are indeed capable of generating effective interaction designs, supporting the overall effectiveness of the approach proposed in Chapter 3.

From these points, I conclude that the dissertation complemented the body of knowledge with a generative concept by showing that the proposed approach to utilize off-the-shelf ML models is contestable, substantive, and defensible. Notably, by retrospectively looking back at the five examples in Table 24, we can confirm the generative power of this concept, as it allows us to imagine further possible interaction designs. For instance, as depicted

Table 24: Summary of the interaction designs presented in this dissertation.

Strategy	Chapter	Domain	User needs	Off-the-shelf models	Interaction design
Strategy #1	Chapter 4	Video-based lectures	To return attention to a video when distracted by side tasks.	Vision-based distraction detection models [298, 79]	Design an intervention based on the human nature of speech communication that functions without consuming conscious awareness.
	Chapter 5	Intellectual tasks	To avoid procrastination and maintain engagement with tasks.	Large generative models [35, 250]	Deliver a context-aware and variational intervention based on the elaboration likelihood model, avoiding the risk of losing trust due to users' high expectations.
	Chapter 6	Photo editing	To edit photos exploratory while incorporating a professional-like style.	(The encoder network of) style transfer models [170, 100]	Utilize parametric design to seamlessly integrate style transfer to professional-like examples into the user's familiar design space.
Strategy #2	Chapter 7	Music composition	To compose music creatively even with limited musical knowledge.	Text-to-audio generation models [177] and (the API of) large language models	Facilitate users' understanding of new vocabulary and its impact on the generation results through iterative exploration.
	Chapter 8	Speech transcription	To transcribe speech that is unclear for ASR models efficiently.	Speech recognition models [101] (and their streaming APIs)	Enable effective collaboration with models by helping users improve their respeaking skills via interaction and feedback.

in Table 25, we can identify some common factors in the user needs that these examples have addressed. From this observation, it can be inferred that, for example, when a user's long-term objectives and immediate desires are in conflict, it would be plausible to consider employing Strategy #1. Of course, it is also possible to apply the proposed approach in scenarios devoid of these factors (*e.g.*, considering the use of off-the-shelf ML models would be effective when user needs are urgent such that they do not allow the luxury of time to prepare tailored ML models). Furthermore, augmenting Strategy #1 by utilizing insights from social psychology, instead of individual cognitive processes, could enable addressing societal needs with off-the-shelf models. These possibilities indicate that accumulating applications of the proposed approach represents an intriguing direction for the future work of this dissertation.

9.2 Ethical considerations

Meanwhile, I am aware of the ethical considerations necessary for the proposed approach. First, as is common in all ML-based systems, clarifying the risks involved and providing transparent explanations and necessary oversight will likely be broadly demanded, in line with requirements like those of the EU AI Act [75]. In addition, particularly for interaction designs that focus on the plasticity of the human side, there is an undeniable potential to inadvertently alter user behavior. From this perspective, obtaining informed consent is crucial upon the deployment of interaction designs based on the proposed approach, and thus, this dissertation has attempted to present interaction designs that function effectively even when subjected to the process of informed consent. Indeed, this dissertation aligns with the shared recognition of the importance of human safety to the discussions around human-centered AI. At the same time, it has also focused on the constraints arising from paying less attention to the plasticity of the human side. Understanding this nuance might be facilitated by embracing the idea of *humanity-centered design*, an idea advocated by Norman [215], who significantly contributed to the proliferation of human-centered design [217] and later pointed out its limitations [216]. Humanity-centered design goes beyond individual user experiences and interactions, considering broader implications for human values, societal impact, and ethical considerations, emphasizing the importance of a holistic view of humanity. In this context, the dissertation has proposed an approach to address individual user needs by utilizing off-the-shelf ML models, with an eye toward the sustainability of the entire ecosystem. This approach embodies the belief that it is possible to balance considerations for individual users with holistic sustainability. However, identifying specific means to achieve this balance remains a part of future work.

At the same time, this point of view necessitates the consideration of sustainability around the community, supporting the creation and sharing of off-the-shelf ML models. As described in Section 2.2, the current community consists of various stakeholders from academia, industry, and start-ups. While there have been some studies on how off-the-shelf models were created and shared within communities like HuggingFace [43, 140], the discussion on how to support such activities in a sustainable manner is still limited. If the use of off-the-shelf models by end-users becomes more widespread, it could potentially contribute back to the community, fostering even more efficient model creation. Discussing these possibilities from a system point of view, with the aim of serving humanity, is indeed

Table 25: Common factors behind the user needs that have been addressed by the presented interaction designs.

Strategy	Chapter	User needs	Factors
Strategy #1	Chapter 4	To return attention to a video when distracted by side tasks.	There is a conflict between a user's long-term objectives and immediate desires.
	Chapter 5	To avoid procrastination and maintain engagement with tasks.	
	Chapter 6	To edit photos exploratory while incorporating a professional-like style.	There is a difficulty in clearly and explicitly defining a user's objectives.
Strategy #2	Chapter 7	To compose music creatively even with limited musical knowledge.	There is a gap between an ML model's expected inputs and a user's capabilities.
	Chapter 8	To transcribe speech that is unclear for ASR models efficiently.	

another avenue for future work.

9.3 Toward grassroots development of interactive systems harnessing off-the-shelf machine learning models

As stated in Chapters 1 and 3, the ultimate goal behind presenting the proposed approach as a generative concept is realizing a world where end-users can develop interactive systems using off-the-shelf ML models to address their open-ended needs in a grassroots manner. This ambition is partly driven by the expectation that the barriers for end-users to develop interactive systems will be lowered through the advancement of ML technologies. Specifically, several methods utilizing off-the-shelf ML models, like Codex [48], have been proposed to assist in programming [375, 255, 171], holding the promise of enabling even beginners to develop a variety of systems [138]. In other words, it becomes conceivable that end-users could consult off-the-shelf models to develop their own interactive systems harnessing other off-the-shelf models. Indeed, during the COVID-19 pandemic, I observed practitioners utilizing off-the-shelf tools to address their unique communication needs in complex situations [341], which has made me believe the plausibility that such grassroots solutions will gain traction. Note that such an idea of enabling end-users to develop or customize systems has been discussed as end-user programming, end-user software engineering, or end-user development [36, 172, 153, 19]. There have also been proposals of toolkits that enable end-users to prototype interactive systems using ML models [334, 74].

This dissertation complements this thread of research from a different layer by offering strategies on how to reconcile off-the-shelf models with user needs on top of such toolkits. Specifically, if the challenges for preparing tailored ML models persist, as mentioned in Section 2.3, the next challenge for end-users would indeed be how to effectively utilize off-the-shelf models to solve individual needs. Then, I expect that the proposed approach provides them with a good starting point, as a generative concept. Furthermore, we can envisage the creation of an ML-based system that serves as a consultative tool for users (e.g., a conversational agent) by harnessing the presented strategies. If this system can effectively guide users in addressing their individual needs using off-the-shelf models, it would lead to the (semi-)automatic generation of effective interaction designs.

At the same time, this dissertation considers the strategies outlined in Section 3.2 not as the only possible means and acknowledges the potential for exploring other strategies. Here, as a discipline closest to understanding humans within computer science, I believe that HCI researchers are uniquely positioned to codify further contributions in making the human side more approachable to off-the-shelf ML models. In this regard, I look forward to future research regarding the utilization of off-the-shelf models that would span from specific instances to general theories.

10 Conclusion

> Our repertoire of food-processing methods altered the genetic selection pressures on our digestive system by gradually supplanting some of its functions with cultural substitutes... The energy savings from the externalization of digestive functions by cultural evolution became one component in a suite of adjustments that permitted our species to build and run bigger and bigger brains. – Joseph Henrich [114]

This dissertation proposed an approach that utilizes the power of interaction design to address open-ended user needs by leveraging off-the-shelf ML models. Its underlying philosophy is to conceptualize the relationship between off-the-shelf models and end-users as a bivariate problem. This yielded strategies for generating interaction designs that make the human side approachable for off-the-shelf ML models. The effectiveness of these strategies was confirmed through five actual examples of interaction designs that were constructed against different tasks across various domains, suggesting that the proposed approach can serve as an actionable guidance or concept to generate new interaction designs.

As mentioned at the beginning of this dissertation, plasticity is one of the special features of the human brain [300, 268]. This means that our knowledge and behavior are influenced by external environments—including tools—which may also affect evolutionary pressure [114]. It would be one of the reasons why a holistic perspective is required in designing interactive systems [215] so as not to unintentionally affect individual users, or even humanity. Simultaneously, this dissertation suggested that there are discoveries to be made by intentionally focusing on human adaptability. With this in mind, I hope to contribute to the realization of a world where humans and machines co-learn, thereby expanding the frontiers of humanity.

References

[1] Alexander Travis Adams, Jean Marcel dos Reis Costa, Malte F. Jung, and Tanzeem Choudhury. "Mindless Computing: Designing Technologies to Subtly Influence Behavior". In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 719–730. DOI: 10.1145/2750858.2805843.

[2] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. "MusicLM: Generating Music From Text". In: *arXiv* 2301.11325 (2023), pp. 1–15. DOI: 10.48550/arXiv.2301.11325.

[3] Taketo Akama. "Connective Fusion: Learning Transformational Joining of Sequences with Application to Melody Creation". In: *Proceedings of the 21th International Society for Music Information Retrieval Conference*. ISMIR, 2020, pp. 46–53.

[4] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. "Optuna: A Next-Generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019, pp. 2623–2631. DOI: 10.1145/3292500.3330701.

[5] Obead Alhadreti and Pam J. Mayhew. "Rethinking Thinking Aloud: A Comparison of Three Think-Aloud Protocols". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 44. DOI: 10.1145/3173574.3173618.

[6] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. "Guidelines for Human-AI Interaction". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 3. ACM, 2019, pp. 1–13. DOI: 10.1145/3290605.3300233.

[7] Kristina Andersen and Peter Knees. "Conversations with Expert Users in Music Retrieval and Research Challenges for Creative MIR". In: *Proceedings of the 17th International Society for Music Information Retrieval Conference*. ISMIR, 2016, pp. 122–128.

[8] Riku Arakawa, Shinnnosuke Takamichi, and Hiroshi Saruwatari. "Implementation of DNN-Based Real-Time Voice Conversion and Its Improvements by Audio Data Augmentation and Mask-Shaped Device". In: *Proceedings of the 10th ISCA Speech Synthesis Workshop*. ISCA, 2019, pp. 93–98. DOI: 10.21437/SSW.2019-17.

[9] Riku Arakawa and Hiromu Yakura. "AI for Human Assessment: What Do Professional Assessors Need?" In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 378. ACM, 2023, pp. 1–7. DOI: 10.1145/3544549.3573849.

[10] Riku Arakawa and Hiromu Yakura. "INWARD: A Computer-Supported Tool for Video-Reflection Improves Efficiency and Effectiveness in Executive Coaching". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 574. ACM, 2020, pp. 1–13. DOI: 10.1145/3313831.3376703.

[11] Riku Arakawa and Hiromu Yakura. “Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 99. ACM, 2021, pp. 1–15. DOI: 10.1145/3411764.3445339.

[12] Riku Arakawa and Hiromu Yakura. “REsCUE: A Framework for REal-Time Feedback on Behavioral CUEs Using Multimodal Anomaly Detection”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 572. ACM, 2019, pp. 1–13. DOI: 10.1145/3290605.3300802.

[13] Riku Arakawa, Hiromu Yakura, and Masataka Goto. “BeParrot: Efficient Interface for Transcribing Unclear Speech via Respeaking”. In: *Proceedings of the 27th ACM International Conference on Intelligent User Interfaces*. ACM, 2022, pp. 832–840. DOI: 10.1145/3490099.3511164.

[14] Riku Arakawa, Hiromu Yakura, and Masataka Goto. “CatAlyst: Domain-Extensible Intervention for Preventing Task Procrastination Using Large Generative Models”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 157. ACM, 2023, pp. 1–19. DOI: 10.1145/3544548.3581133.

[15] Ryan Shaun J. D. Baker, Sidney K. D’Mello, Ma. Mercedes T. Rodrigo, and Arthur C. Graesser. “Better to Be Frustrated Than Bored: The Incidence, Persistence, and Impact of Learners’ Cognitive–Affective States During Interactions with Three Different Computer-Based Learning Environments”. In: *International Journal of Human-Computer Studies* 68.4 (2010), pp. 223–241. DOI: 10.1016/j.ijhcs.2009.12.003.

[16] Hyunseung Bang and Daniel Selva. “iFEED: Interactive Feature Extraction for Engineering Design”. In: *Proceedings of the 28th International Conference on Design Theory and Methodology*. 37. American Society of Mechanical Engineers, 2016, pp. 1–11. DOI: 10.1115/detc2016-60077.

[17] Jeffrey Bardzell. “Interaction Criticism: An Introduction to the Practice”. In: *Interacting with Computers* 23.6 (2011), pp. 604–621. DOI: 10.1016/J.INTCOM.2011.07.001.

[18] Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. “Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production”. In: *Speech Communication* 33.1–2 (2001), pp. 5–22. DOI: 10.1016/S0167-6393(00)00067-4.

[19] Barbara Rita Barricelli, Fabio Cassano, Daniela Fogli, and Antonio Piccinno. “End-User Development, End-User Programming and End-User Software Engineering: A Systematic Mapping Study”. In: *Journal of Systems and Software* 149 (2019), pp. 101–137. DOI: 10.1016/J.JSS.2018.11.041.

[20] Meltem Huri Baturay. “An Overview of the World of MOOCs”. In: *Procedia – Social and Behavioral Sciences* 174 (2015), pp. 427–433. DOI: 10.1016/j.sbspro.2015.01.685.

[21] Maude Beauchemin, Louis De Beaumont, Phetsamone Vannasing, Aline Turcotte, Claudine Arcand, Pascal Belin, and Maryse Lassonde. “Electrophysiological Markers of Voice Familiarity”. In: *The European Journal of Neuroscience* 23.11 (2006), pp. 3081–3086. DOI: 10.1111/j.1460-9568.2006.04856.x.

[22] Malikeh Beheshtifar, Hadis Hoseinifar, and Moghadam Nekoie Moghadam. "Effect Procrastination on Work-Related Stress". In: *European Journal of Economics, Finance and Administrative Sciences* 38.38 (2011), pp. 59–64.

[23] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2021, pp. 610–623. DOI: 10.1145/3442188.3445922.

[24] Daniel E Berlyne. *Conflict, Arousal, and Curiosity*. New York, NY: McGraw-Hill, 1960.

[25] Angela M. Zavaleta Bernuy, Ziwen Han, Hammad Shaikh, Qi Yin Zheng, Lisa-Angelique Lim, Anna N. Rafferty, Andrew Petersen, and Joseph Jay Williams. "How Can Email Interventions Increase Students' Completion of Online Homework? A Case Study Using A/B Comparisons". In: *Proceedings of the 12th International Learning Analytics and Knowledge Conference*. ACM, 2022, pp. 107–118. DOI: 10.1145/3506860.3506874.

[26] Angela M. Zavaleta Bernuy, Qi Yin Zheng, Hammad Shaikh, Andrew Petersen, and Joseph Jay Williams. "Investigating the Impact of Online Homework Reminders Using Randomized A/B Comparisons". In: *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. ACM, 2021, pp. 921–927. DOI: 10.1145/3408877.3432427.

[27] Jonathan Bidwell and Henry Fuchs. *Classroom Analytics: Measuring Student Engagement with Automated Gaze Tracking*. Tech. rep. Chapel Hill, NC: Department of Computer, University of North Carolina at Chapel Hill, 2011, pp. 1–17. DOI: 10.13140/RG.2.1.4865.6242.

[28] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[29] Christopher M. Bishop and Hugh Bishop. *Deep Learning – Foundations and Concepts*. Berlin, Germany: Springer, 2024. DOI: 10.1007/978-3-031-45468-4.

[30] Julian O. Blosiu. "Use of Synectics as an Idea Seeding Technique to Enhance Design Creativity". In: *Proceedings of the 1999 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 1999, pp. 1001–1006. DOI: 10.1109/icsmc.1999.823365.

[31] Sara A. Bly and Elizabeth F. Churchill. "Design Through Matchmaking: Technology in Search of Users". In: *Interactions* 6.2 (1999), pp. 23–31. DOI: 10.1145/296165.296174.

[32] Jean-Pierre Briot and François Pachet. "Deep Learning for Music Generation: Challenges and Directions". In: *Neural Computing and Applications* 32.4 (2020), pp. 981–993. DOI: 10.1007/s00521-018-3813-6.

[33] Eric Brochu, Nando de Freitas, and Abhijeet Ghosh. "Active Preference Learning with Discrete Choice Data". In: *Proceedings of the 21st Annual Conference on Neural Information Processing Systems*. Neural Information Processing Systems Foundation, 2007, pp. 409–416.

[34] John Brooke. “SUS: A ‘Quick and Dirty’ Usability Scale”. In: *Usability Evaluation In Industry*. CRC Press, 1996, pp. 207–212.

[35] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. “Language Models Are Few-Shot Learners”. In: *Proceedings of the 2020 Annual Conference on Neural Information Processing Systems*. Curran Associates, 2020, pp. 1877–1901.

[36] Margaret M. Burnett, Curtis R. Cook, and Gregg Rothermel. “End-User Software Engineering”. In: *Communications of the ACM* 47.9 (2004), pp. 53–58. DOI: 10.1145/1015864.1015889.

[37] James C. Byers, Alvah C. Bittner, and Susan G. Hill. “Traditional and Raw Task Load Index (TLX) Correlations: Are Paired Comparisons Necessary?” In: *Proceedings of the 1989 Annual International Industrial Ergonomics and Safety Conference*. Taylor & Francis, 1989, pp. 481–485.

[38] Antoine Caillon and Philippe Esling. “RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis”. In: *arXiv* 2111.05011 (2021), pp. 1–15. DOI: 10.48550/arXiv.2111.05011.

[39] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. “Extracting Training Data From Large Language Models”. In: *Proceedings of the 30th USENIX Security Symposium*. USENIX Association, 2021, pp. 2633–2650.

[40] Filippo Carnovalini and Antonio Rodà. “Computational Creativity and Music Generation Systems: An Introduction to the State of the Art”. In: *Frontiers in Artificial Intelligence* 3 (2020), p. 14. DOI: 10.3389/frai.2020.00014.

[41] John M. Carroll and Wendy A. Kellogg. “Artifact as Theory-Nexus: Hermeneutics Meets Theory-Based Design”. In: *Proceedings of the 1989 CHI Conference on Human Factors in Computing Systems*. ACM, 1989, pp. 7–14. DOI: 10.1145/67449.67452.

[42] John M. Carroll and Judith Reitman Olson. “Mental Models in Human-Computer Interaction”. In: *Handbook of Human-Computer Interaction*. North-Holland, 1988, pp. 45–65. DOI: 10.1016/B978-0-444-70536-5.50007-5.

[43] Joel Castaño, Silverio Martínez-Fernández, Xavier Franch, and Justus Bogner. “Analyzing the Evolution and Maintenance of ML Models on Hugging Face”. In: *arXiv* 2311.13380 (2023), pp. 1–11. DOI: 10.48550/ARXIV.2311.13380.

[44] Kemal Aydin M. Emre Celebi, ed. *Unsupervised Learning Algorithms*. Cham, Switzerland: Springer International Publishing, 2016. DOI: 10.1007/978-3-319-24211-8.

[45] Haiwen Chen, Jane Epstein, and Emily Stern. "Neural Plasticity After Acquired Brain Injury: Evidence From Functional Neuroimaging". In: *PM&R* 2.12S (2010), S306–S312. DOI: 10.1016/j.pmrj.2010.10.006.

[46] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. "BeautyGlow: On-Demand Makeup Transfer Framework with Reversible Generative Network". In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 10042–10050. DOI: 10.1109/CVPR.2019.01028.

[47] Ke Chen, Cheng-i Wang, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. "Music SketchNet: Controllable Music Generation via Factorized Representations of Pitch and Rhythm". In: *Proceedings of the 21st International Society for Music Information Retrieval Conference*. ISMIR, 2020, pp. 77–84.

[48] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. "Evaluating Large Language Models Trained on Code". In: *arXiv* 2107.03374 (2021), pp. 1–35. DOI: 10.48550/arXiv.2107.03374.

[49] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. "Gmail Smart Compose: Real-Time Assisted Writing". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019, pp. 2287–2295. DOI: 10.1145/3292500.3330723.

[50] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[51] Mark H. Chignell, Lu Wang, Atefeh Zare, and Jamy Li. "The Evolution of HCI and Human Factors: Integrating Human and Artificial Intelligence". In: *ACM Transactions on Computer-Human Interaction* 30.2 (2023), pp. 1–30. DOI: 10.1145/3557891.

[52] Hyunsung Cho, Daeun Choi, Donghwi Kim, Wan Ju Kang, Eun Kyoung Choe, and Sung-Ju Lee. "Reflect, Not Regret: Understanding Regretful Smartphone Use with App Feature-Level Analysis". In: *Proceedings of the ACM on Human-Computer Interaction* 5.456 (2021), pp. 1–36. DOI: 10.1145/3479600. URL: https://doi.org/10.1145/3479600.

[53] Jaemin Cho, Abhay Zala, and Mohit Bansal. "DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers". In: *arXiv* 2202.04053 (2022), pp. 1–20. DOI: 10.48550/arXiv.2202.04053.

[54] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. "PaLM: Scaling Language Modeling with Pathways". In: *Journal of Machine Learning Research* 24.240 (2023), pp. 1–113.

[55] Francesco Cirillo. *The Pomodoro Technique: The Life-Changing Time-Management System*. New York, NY: Random House, 2018.

[56] James Clear. *The Chemistry of Building Better Habits*. https://jamesclear.com/chemistry-habits. Last accessed on September 2022. 2015.

[57] Tatiana Conde, Óscar F. Gonçalves, and Ana P. Pinheiro. "Paying Attention to My Voice or Yours: An ERP Study with Words". In: *Biological Psychology* 111 (2015), pp. 40–52. DOI: 10.1016/j.biopsycho.2015.07.014.

[58] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. "Unsupervised Cross-Lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

[59] Crispin Coombs. "Will COVID-19 Be the Tipping Point for the Intelligent Automation of Work? A Review of the Debate and Implications for Research". In: *International Journal of Information Management* 102182 (2020), pp. 1–4. DOI: 10.1016/j.ijinfomgt.2020.102182.

[60] Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. "Gaze Tutor: A Gaze-Reactive Intelligent Tutoring System". In: *International Journal of Human-Computer Studies* 70.5 (2012), pp. 377–398. DOI: 10.1016/j.ijhcs.2012.01.004.

[61] Shuqi Dai, Zeyu Jin, Celso Gomes, and Roger B. Dannenberg. "Controllable Deep Melody Generation via Hierarchical Music Structure Representation". In: *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021, pp. 143–150.

[62] Salvador Dalí. *Dali by Dali*. New York, NY: Abrams, 1970.

[63] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. "Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries". In: *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 98. ACM, 2022, pp. 1–13. DOI: 10.1145/3526113.3545672.

[64] Fred D. Davis. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology". In: *MIS Quarterly* 13.3 (1989), pp. 319–340. DOI: 10.2307/249008.

[65] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 509. ACM, 2020, pp. 1–12. DOI: 10.1145/3313831.3376638.

[66] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild". In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 5202–5211. DOI: 10.1109/CVPR42600.2020.00525.

[67] Emmanuel Deruty, Maarten Grachten, Stefan Lattner, Javier Nistal, and Cyran Aouameur. "On the Development and Practice of AI Technology for Contemporary Popular Music Production". In: *Transactions of the International Society for Music Information Retrieval* 5.1 (2022), p. 35. DOI: 10.5334/tismir.100.

[68] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.

[69] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. "Jukebox: A Generative Model for Music". In: *arXiv* 2005.00341 (2020), pp. 1–20. DOI: 10.48550/arXiv.2005.00341.

[70] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err". In: *Journal of Experimental Psychology: General* 144.1 (2015), pp. 114–126. DOI: 10.1037/xge0000033.

[71] Carl Doersch. "Tutorial on Variational Autoencoders". In: *arXiv* 1606.05908 (2016), pp. 1–23.

[72] Hao-Wen Dong and Yi-Hsuan Yang. "Convolutional Generative Adversarial Networks with Binary Neurons for Polyphonic Music Generation". In: *Proceedings of the 19th International Society for Music Information Retrieval Conference*. ISMIR, 2018, pp. 190–196.

[73] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. "Clotho: an Audio Captioning Dataset". In: *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 736–740. DOI: 10.1109/ICASSP40776.2020.9052990.

[74] Ruofei Du, Na Li, Jing Jin, Michelle Carney, Scott Miles, Maria Kleiner, Xiuxiu Yuan, Yinda Zhang, Anuva Kulkarni, Xingyu Liu, Ahmed Sabie, Sergio Orts-Escolano, Abhishek Kar, Ping Yu, Ram Iyengar, Adarsh Kowdle, and Alex Olwal. "Rapsai: Accelerating Machine Learning Prototyping of Multimedia Applications Through Visual Programming". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 125. ACM, 2023, pp. 1–23. DOI: 10.1145/3544548.3581338.

[75] Lilian Edwards. *Expert Explainer: The EU AI Act Proposal*. Tech. rep. London, UK: Ada Lovelace Institute, 2022.

[76] Mennatallah El-Assady and Caterina Moruzzi. "Which Biases and Reasoning Pitfalls Do Explanations Trigger? Decomposing Communication Processes in Human-AI Interaction". In: *IEEE Computer Graphics and Applications* 42.6 (2022), pp. 11–23. DOI: 10.1109/MCG.2022.3200328.

[77] Douglas C. Engelbart. *Augmenting Human Intellect: A Conceptual Framework*. Tech. rep. AFOSR-3223. Menlo Park, US: Stanford Research Institute, 1962.

[78] Daniel Fallman. "Design-Oriented Human-Computer Interaction". In: *Proceedings of the 2003 CHI Conference on Human Factors in Computing Systems*. ACM, 2003, pp. 225–232. DOI: 10.1145/642611.642652.

[79] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. "RMPE: Regional Multi-Person Pose Estimation". In: *Proceedings of the 16th IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 2353–2362. DOI: 10.1109/ICCV.2017.256.

[80] Jose D. Fernández and Francisco J. Vico. "AI Methods in Algorithmic Composition: A Comprehensive Survey". In: *Journal of Artificial Intelligence Research* 48 (2013), pp. 513–582. DOI: 10.1613/jair.3908.

[81] Robert S. Fish, Robert E. Kraut, Robert W. Root, and Ronald E. Rice. "Evaluating Video as a Technology for Informal Communication". In: *Proceedings of the 1992 CHI Conference on Human Factors in Computing Systems*. ACM, 1992, pp. 37–48. DOI: 10.1145/142750.142755.

[82] B. J. Fogg. "A Behavior Model for Persuasive Design". In: *Proceedings of the 4th International Conference on Persuasive Technology*. 40. ACM, 2009, pp. 1–7. DOI: 10.1145/1541948.1541999.

[83] Michael D. Frakes and Melissa F. Wasserman. *Procrastination in the Workplace: Evidence From the US Patent Office*. Tech. rep. National Bureau of Economic Research, 2016.

[84] Giorgio Franceschelli and Mirco Musolesi. "Creativity and Machine Learning: A Survey". In: *arXiv* 2104.02726 (2021), pp. 1–35. DOI: 10.48550/arXiv.2104.02726.

[85] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677.

[86] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. "Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors". In: *Proceedings of the 17th European Conference on Computer Vision*. Springer, 2022, pp. 89–106. DOI: 10.1007/978-3-031-19784-0_6.

[87] Wensheng Gan, Shicheng Wan, and Philip S. Yu. "Model-as-a-Service (MaaS): A Survey". In: *arXiv* 2311.05804 (2023), pp. 1–11. DOI: 10.48550/arxiv.2311.05804.

[88] Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham. "The Effects of Automatic Speech Recognition Quality on Human Transcription Latency". In: *Proceedings of the 13th Web for All Conference*. 23. ACM, 2016, pp. 1–8. DOI: 10.1145/2899475.2899478.

[89] William W. Gaver. "What Should We Expect From Research Through Design?" In: *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 937–946. DOI: 10.1145/2207676.2208538.

[90] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events". In: *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.

[91] John S. Gero. "Creativity, Emergence and Evolution in Design". In: *Knowledge-Based Systems* 9.7 (1996), pp. 435–448. DOI: 10.1016/S0950-7051(96)01054-4.

[92] Anne-Sophie Ghyselen, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen, and Arjan van Hessen. "Clearing the Transcription Hurdle in Dialect Corpus Building: The Corpus of Southern Dutch Dialects as Case Study". In: *Frontiers in Artificial Intelligence* 3 (2020), p. 10. DOI: 10.3389/frai.2020.00010.

[93] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment". In: *JMIR Medical Education* 9 (2023), e45312. DOI: 10.2196/45312.

[94] Karen Goldschmidt. "The COVID-19 Pandemic: Technology Use to Support the Wellbeing of Children". In: *Journal of Pediatric Nursing* 53 (2020), pp. 88–90. DOI: 10.1016/j.pedn.2020.04.013.

[95] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. "Generative Adversarial Networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144. DOI: 10.1145/3422622.

[96] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *Proceedings of the 3rd International Conference on Learning Representations*. OpenReview.net, 2015.

[97] Joyce Eastlund Gromko. "Perceptual Differences Between Expert and Novice Music Listeners: A Multidimensional Scaling Analysis". In: *Psychology of Music* 21.1 (1993), pp. 34–47. DOI: 10.1177/030573569302100103.

[98] Tovi Grossman, George W. Fitzmaurice, and Ramtin Attar. "A Survey of Software Learnability: Metrics, Methodologies and Guidelines". In: *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 649–658. DOI: 10.1145/1518701.1518803.

[99] Jonathan Grudin. "Three Faces of Human-Computer Interaction". In: *IEEE Annals of the History of Computing* 27.4 (2005), pp. 46–62. DOI: 10.1109/MAHC.2005.67.

[100] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. "LADN: Local Adversarial Disentangling Network for Facial Makeup and De-Makeup". In: *Proceedings of the 17th IEEE International Conference on Computer Vision*. IEEE, 2019, pp. 10480–10489. DOI: 10.1109/ICCV.2019.01058.

[101] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. "Conformer: Convolution-Augmented Transformer for Speech Recognition". In: *Proceedings of the 21st Annual Conference of the International Speech Communication Association*. ISCA, 2020, pp. 5036–5040. DOI: 10.21437/Interspeech.2020-3015.

[102] Jingtao Guo, Zhenzhen Qian, Zuowei Zhou, and Yi Liu. "MulGAN: Facial Attribute Editing by Exemplar". In: *arXiv* 1912.12396 (2019), pp. 1–14.

[103] Linsong Guo, Qin Zhang, and Shuxia Han. "Agricultural Machinery Safety Alert System Using Ultrasonic Sensors". In: *Journal of Agricultural Safety and Health* 8.4 (2002), pp. 385–396. DOI: 10.13031/2013.10219.

[104] Philip J. Guo, Juho Kim, and Rob Rubin. "How Video Production Affects Student Engagement: an Empirical Study of MOOC Videos". In: *Proceedings of the 1st ACM Conference on Learning @ Scale*. ACM, 2014, pp. 41–50. DOI: 10.1145/2556325.2566239.

[105] Reinhold Haeb-Umbach, Jahn Heymann, Lukas Drude, Shinji Watanabe, Marc Delcroix, and Tomohiro Nakatani. "Far-Field Automatic Speech Recognition". In: *Proceedings of the IEEE* 109.2 (2021), pp. 124–148. DOI: 10.1109/JPROC.2020.3018668.

[106] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. "Pre-Trained Models: Past, Present and Future". In: *AI Open* 2 (2021), pp. 225–250. DOI: 10.1016/J.AIOPEN.2021.08.002.

[107] Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Advances in Psychology* 52 (1988), pp. 139–183. DOI: 10.1016/s0166-4115(08)62386-9.

[108] Melanie Hartmann. "Challenges in Developing User-Adaptive Intelligent User Interfaces". In: *Proceedings of the 17th Workshop on Adaptivity and User Modeling in Interactive Systems*. Technische Universität Darmstadt, 2009, pp. 6–10.

[109] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving Deep Into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *Proceedings of the 15th IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 1026–1034. DOI: 10.1109/ICCV.2015.123.

[110] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. "AttGAN: Facial Attribute Editing by Only Changing What You Want". In: *IEEE Transactions on Image Processing* 28.11 (2019), pp. 5464–5478. DOI: 10.1109/TIP.2019.2916751.

[111] Jeffrey Heer. "Agency Plus Automation: Designing Artificial Intelligence Into Interactive Systems". In: *Proceedings of the National Academy of Sciences* 116.6 (2019), pp. 1844–1850. DOI: 10.1073/pnas.1807184115.

[112] William E. Hefley and Dianne Murray. "Intelligent User Interfaces". In: *Proceedings of the 1st ACM International Workshop on Intelligent User Interfaces*. ACM, 1993, pp. 3–10. DOI: 10.1145/169891.169892.

[113] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning". In: *Journal of Machine Learning Research* 21.248 (2020), pp. 1–43.

[114] Joseph Henrich. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, US: Princeton University Press, 2015. DOI: 10.2307/j.ctvc77f0d.

[115] Daniel Holden, Ikhsanul Habibie, Ikuo Kusajima, and Taku Komura. "Fast Neural Style Transfer for Motion Data". In: *IEEE Computer Graphics and Applications* 37.4 (2017), pp. 42–49. DOI: 10.1109/MCG.2017.3271464.

[116] Shinichi Homma, Akio Kobayashi, Takahiro Oku, Shoei Sato, Toru Imai, and Tohru Takagi. "New Real-Time Closed-Captioning System for Japanese Broadcast News Programs". In: *Proceedings of the 11th International Conference on Computers Helping People with Special Needs*. Springer Berlin Heidelberg, 2008, pp. 651–654. DOI: 10.1007/978-3-540-70540-6_93.

[117] Richang Hong, Meng Wang, Mengdi Xu, Shuicheng Yan, and Tat-Seng Chua. "Dynamic Captioning: Video Accessibility Enhancement for Hearing Impairment". In: *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 2010, pp. 421–430. DOI: 10.1145/1873951.1874013.

[118] Kristina Höök. "Steps to Take Before Intelligent User Interfaces Become Real". In: *Interacting with Computers* 12.4 (2000), pp. 409–426. DOI: 10.1016/S0953-5438(99)00006-5.

[119] Kristina Höök and Jonas Löwgren. "Strong Concepts: Intermediate-Level Knowledge in Interaction Design Research". In: *ACM Transactions on Computer-Human Interaction* 19.23 (2012), pp. 1–18. DOI: 10.1145/2362364.2362371.

[120] Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. "Advances in Joint CTC-Attention Based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM". In: *Proceedings of the 18th Annual Conference of the International Speech Communication Association*. ISCA, 2017, pp. 949–953.

[121] Kurt Hornik, Maxwell B. Stinchcombe, and Halbert White. "Multilayer Feedforward Networks Are Universal Approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.

[122] Eric Horvitz. "Principles of Mixed-Initiative User Interfaces". In: *Proceedings of the 1999 CHI Conference on Human Factors in Computing Systems*. ACM, 1999, pp. 159–166. DOI: 10.1145/302979.303030.

[123] Lauren C. Howe and Jochen I. Menges. "Remote Work Mindsets Predict Emotions and Productivity in Home Office: A Longitudinal Study of Knowledge Workers During the COVID-19 Pandemic". In: *Human–Computer Interaction* in press (2021), pp. 1–27. DOI: 10.1080/07370024.2021.1987238.

[124] Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. "Compound Word Transformer: Learning to Compose Full-Song Music Over Dynamic Directed Hypergraphs". In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. AAAI Press, 2021, pp. 178–186. DOI: 10.1609/AAAI.V35I1.16091.

[125] Yiwei Hu, Julie Dorsey, and Holly E. Rushmeier. "A Novel Framework for Inverse Procedural Texture Modeling". In: *ACM Transactions on Graphics* 38.186 (2019), pp. 1–14. DOI: 10.1145/3355089.3356516.

[126] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. "Exposure: A White-Box Photo Post-Processing Framework". In: *ACM Transactions on Graphics* 37.26 (2018), pp. 1–17. DOI: 10.1145/3181974.

[127] Cheng-Zhi Anna Huang, Curtis Hawthorne, Adam Roberts, Monica Dinculescu, James Wexler, Leon Hong, and Jacob Howcroft. "The Bach Doodle: Approachable Music Composition with Machine Learning at Scale". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference*. ISMIR, 2019, pp. 793–800.

[128] Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie Cai. "Human-AI Co-Creation in Songwriting". In: *Proceedings of the 21th International Society for Music Information Retrieval Conference*. ISMIR, 2020, pp. 708–716.

[129] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. "Music Transformer: Generating Music with Long-Term Structure". In: *Proceedings of the 7th International Conference on Learning Representations*. OpenReview.net, 2019.

[130] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. "Multimodal Unsupervised Image-to-Image Translation". In: *Proceedings of the 15th European Conference on Computer Vision*. Springer, 2018, pp. 179–196. DOI: 10.1007/978-3-030-01219-9_11.

[131] Yu-Siang Huang and Yi-Hsuan Yang. "Pop Music Transformer: Beat-Based Modeling and Generation of Expressive Pop Piano Compositions". In: *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020, pp. 1180–1188. DOI: 10.1145/3394171.3413671.

[132] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. "EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-Based Music Generation". In: *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021, pp. 318–325.

[133] Tun-Min Hung, Bo-Yu Chen, Yen-Tung Yeh, and Yi-Hsuan Yang. "A Benchmarking Initiative for Audio-Domain Music Generation Using the FreeSound Loop Dataset". In: *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021, pp. 310–317.

[134] Stephen Hutt, Caitlin Mills, Nigel Bosch, Kristina Krasich, James R. Brockmole, and Sidney K. D'Mello. ""Out of the Fr-Eye-Ing Pan": Towards Gaze-Based Models of Attention During Learning with Technology in the Classroom". In: *Proceedings of the 25th ACM International Conference on User Modeling, Adaptation and Personalization*. ACM, 2017, pp. 94–103. DOI: 10.1145/3079628.3079669.

[135] Toru Imai, Atsushi Matsui, Shinichi Homma, Takeshi Kobayakawa, Kazuo Onoe, Shoei Sato, and Akio Ando. "Speech Recognition with a Re-Speak Method for Subtitling Live Broadcasts". In: *Proceedings of the 7th International Conference on Spoken Language Processing*. ISCA, 2002.

[136] Toastmasters International. *Your Speaking Voice: Tips for Adding Strength and Authority to Your Speaking Voice*. https://www.toastmasters.org/resources/your-speaking-voice. Accessed: August 19, 2019. July 2011.

[137] V. López Jaquero, Francisco Montero, José Pascual Molina, and Pascual González. "Intelligent User Interfaces: Past, Present and Future". In: *Engineering the User Interface*. Springer London, 2008, pp. 1–12. DOI: 10.1007/978-1-84800-136-7_18.

[138] Dhanya Jayagopal, Justin Lubin, and Sarah E. Chasins. "Exploring the Learnability of Program Synthesizers by Novice Programmers". In: *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 64. ACM, 2022, pp. 1–15. DOI: 10.1145/3526113.3545659.

[139] Shulei Ji, Jing Luo, and Xinyu Yang. "A Comprehensive Survey on Deep Music Generation: Multi-Level Representations, Algorithms, Evaluations, and Future Directions". In: *arXiv* 2011.06801 (2020), pp. 1–96. DOI: 10.48550/arxiv.2011.06801.

[140] Wenxin Jiang, Nicholas Synovic, Matt Hyatt, Taylor R. Schorlemmer, Rohan Sethi, Yung-Hsiang Lu, George K. Thiruvathukal, and James C. Davis. "An Empirical Study of Pre-Trained Model Reuse in the Hugging Face Deep Learning Model Registry". In: *Proceedings of the 45th IEEE/ACM International Conference on Software Engineering*. IEEE, 2023, pp. 2463–2475. DOI: 10.1109/ICSE48619.2023.00206.

[141] Wenxin Jiang, Nicholas Synovic, Rohan Sethi, Aryan Indarapu, Matt Hyatt, Taylor R. Schorlemmer, George K. Thiruvathukal, and James C. Davis. "An Empirical Study of Artifacts and Security Risks in the Pre-Trained Model Supply Chain". In: *Proceedings of the 2022 ACM Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses*. ACM, 2022, pp. 105–114. DOI: 10.1145/3560835.3564547.

[142] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, and Mingli Song. "Neural Style Transfer: A Review". In: *IEEE Transactions on Visualization and Computer Graphics* 26.11 (2020), pp. 3365–3385. DOI: 10.1109/TVCG.2019.2921336.

[143] Mayur P. Joshi, Ning Su, Robert D. Austin, and Anand K. Sundaram. "Why So Many Data Science Projects Fail to Deliver". In: *MIT Sloan Management Review* 62 (3 2021), pp. 1–5.

[144] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. "Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion". In: *Proceedings of the 28th European Conference on Information Systems*. AIS, 2020, pp. 1–16.

[145] Daniel Kahneman. *Thinking, Fast and Slow*. New York, US: Farrar, Straus and Giroux, 2011.

[146] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling Laws for Neural Language Models". In: *arXiv* 2001.08361 (2020), pp. 1–19. DOI: 10.48550/arXiv.2001.08361.

[147] Seita Kayukawa, Keita Higuchi, João Guerreiro, Shigeo Morishima, Yoichi Sato, Kris Kitani, and Chieko Asakawa. "BBeep: A Sonic Collision Avoidance System for Blind Travellers and Nearby Pedestrians". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 52. ACM, 2019, pp. 1–12. DOI: 10.1145/3290605.3300282.

[148] Michael Kerres. "Against All Odds: Education in Germany Coping with COVID-19". In: *Postdigital Science and Education* 2.3 (2020), pp. 690–694. DOI: 10.1007/s42438-020-00130-7.

[149] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. "AudioCaps: Generating Captions for Audios in The Wild". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 2019, pp. 119–132. DOI: 10.18653/v1/n19-1011.

[150] Jingoog Kim, Mary Lou Maher, and Safat Siddiqui. "Studying the Impact of AI-Based Inspiration on Human Ideation in a Co-Creative Design System". In: *Proceedings of the 2rd ACM IUI Workshops on Human-AI Co-Creation with Generative Models*. CEUR-WS.org, 2021.

[151] Diederik P. Kingma and Prafulla Dhariwal. "Glow: Generative Flow with Invertible 1x1 Convolutions". In: *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems*. Neural Information Processing Systems Foundation, 2018, pp. 10236–10245.

[152] Jeffrey A. Kleim and Theresa A. Jones. "Principles of Experience-Dependent Neural Plasticity: Implications for Rehabilitation After Brain Damage". In: *Journal of Speech, Language, and Hearing Research* 51.1 (2008), S225–S239. DOI: 10.1044/1092-4388(2008/018).

[153] Amy J. Ko, Robin Abraham, Laura Beckwith, Alan F. Blackwell, Margaret M. Burnett, Martin Erwig, Christopher Scaffidi, Joseph Lawrance, Henry Lieberman, Brad A. Myers, Mary Beth Rosson, Gregg Rothermel, Mary Shaw, and Susan Wiedenbeck. "The State of the Art in End-User Software Engineering". In: *ACM Computing Surveys* 43 (3 2011), pp. 1–44. DOI: 10.1145/1922649.1922658.

[154] Yasuyuki Kobayashi, Maki Ishibashi, and Hitomi Kobayashi. "How Will "Democratization of Artificial Intelligence" Change the Future of Radiologists?" In: *Japanese Journal of Radiology* 37.1 (2018), pp. 9–14. DOI: 10.1007/s11604-018-0793-5.

[155] Mitsuru Kodama. "Digitally Transforming Work Styles in an Era of Infectious Disease". In: *International Journal of Information Management* 102172 (2020), pp. 1–6. DOI: 10.1016/j.ijinfomgt.2020.102172.

[156] Geza Kovacs, Zhengxuan Wu, and Michael S. Bernstein. "Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition". In: *Proceedings of the ACM on Human-Computer Interaction* 2.95 (2018), pp. 1–25. DOI: 10.1145/3274364.

[157] Yuki Koyama and Takeo Igarashi. "Computational Design with Crowds". In: *Computational Interaction*. Oxford University Press, 2018, pp. 153–184. DOI: 10.1093/oso/9780198799603.003.0007.

[158] Yuki Koyama, Issei Sato, and Masataka Goto. "Sequential Gallery for Interactive Visual Design Optimization". In: *ACM Transactions on Graphics* 39.88 (2020), pp. 1–12. DOI: 10.1145/3386569.3392444.

[159] Line Kühnel, Tom Fletcher, Sarang C. Joshi, and Stefan Sommer. "Latent Space Non-Linear Statistics". In: *arXiv* 1805.07632 (2018), pp. 1–9. DOI: 10.48550/arXiv.1805.07632.

[160] Anastasia Kuzminykh and Sean Rintel. "Classification of Functional Attention in Video Meetings". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 419. ACM, 2020, pp. 1–13. DOI: 10.1145/3313831.3376546.

[161] Anastasia Kuzminykh and Sean Rintel. "Low Engagement As a Deliberate Practice of Remote Participants in Video Meetings". In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 321. ACM, 2020, pp. 1–9. DOI: 10.1145/3334480.3383080.

[162] Oh-Wook Kwon and Jun Park. "Korean Large Vocabulary Continuous Speech Recognition with Morpheme-Based Recognition Units". In: *Speech Communication* 39.3–4 (2003), pp. 287–300. DOI: 10.1016/S0167-6393(02)00031-6.

[163] Clarry H. Lay. "At Last, My Research Article on Procrastination". In: *Journal of Research in Personality* 20.4 (1986), pp. 474–495. DOI: 10.1016/0092-6566(86)90127-3.

[164] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. "Diverse Image-to-Image Translation via Disentangled Representations". In: *Proceedings of the 15th European Conference on Computer Vision*. Springer, 2018, pp. 36–52. DOI: 10.1007/978-3-030-01246-5_3.

[165] Ju Hyun Lee, Ning Gu, and Anthony P. Williams. "Parametric Design Strategies for the Generation of Creative Designs". In: *International Journal of Architectural Computing* 12.3 (2014), pp. 263–282. DOI: 10.1260/1478-0771.12.3.263.

[166] Mina Lee, Percy Liang, and Qian Yang. "CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 388. ACM, 2022, pp. 1–19. DOI: 10.1145/3491102.3502030.

[167] Seungyon Claire Lee and Thad Starner. "BuzzWear: Alert Perception in Wearable Tactile Displays on the Wrist". In: *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 433–442. DOI: 10.1145/1753326.1753392.

[168] Clayton Lewis. "A Model of Mental Model Construction". In: *Proceedings of the 1986 CHI Conference on Human Factors in Computing Systems*. ACM, 1986, pp. 306–313. DOI: 10.1145/22627.22388.

[169] Da-Wei Li, Danqing Huang, Tingting Ma, and Chin-Yew Lin. "Towards Topic-Aware Slide Generation For Academic Papers with Unsupervised Mutual Learning". In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. AAAI Press, 2021, pp. 13243–13251. DOI: 10.1609/AAAI.V35I15.17564.

[170] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. "Learning Linear Transformations for Fast Image and Video Style Transfer". In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 3809–3817. DOI: 10.1109/CVPR.2019.00393.

[171] Jenny T. Liang, Chenyang Yang, and Brad A. Myers. "A Large-Scale Survey on the Usability of AI Programming Assistants: Successes and Challenges". In: *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. IEEE, 2024, pp. 605–617.

[172] Henry Lieberman, Fabio Paternò, Markus Klann, and Volker Wulf. "End-User Development: An Emerging Paradigm". In: *End User Development*. Springer, 2006, pp. 1–8. DOI: 10.1007/1-4020-5386-X_1.

[173] Yuyu Lin, Jiahao Guo, Yang Chen, Cheng Yao, and Fangtian Ying. "It Is Your Turn: Collaborative Ideation with a Co-Creative Robot Through Sketch". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 1–14. DOI: 10.1145/3313831.3376258.

[174] Erin Chao Ling, Iis Tussyadiah, Aarni Tuomi, Jason Stienmetz, and Athina Ioannou. "Factors Influencing Users' Adoption and Use of Conversational Agents: A Systematic Review". In: *Psychology & Marketing* 38.7 (2021), pp. 1031–1051. DOI: 10.1002/mar.21491.

[175] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. "How WEIRD Is CHI?" In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 143. ACM, 2021, p. 14. DOI: 10.1145/3411764.3445488.

[176] Chien-Hung Liu and Chuan-Kang Ting. "Computational Intelligence in Music Composition: A Survey". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 1.1 (2017), pp. 2–15. DOI: 10.1109/TETCI.2016.2642200.

[177] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models". In: *Proceedings of the 2023 International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 21450–21474.

[178] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. "Unsupervised Image-to-Image Translation Networks". In: *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*. Neural Information Processing Systems Foundation, 2017, pp. 700–708.

[179] Peng Liu and Frank K. Soong. "Word Graph Based Speech Recognition Error Correction by Handwriting Input". In: *Proceedings of the 8th ACM International Conference on Multimodal Interfaces*. ACM, 2006, pp. 339–346. DOI: 10.1145/1180995.1181059.

[180] Yikun Liu, Yuan Jia, Wei Pan, and Mark S. Pfaff. "Supporting Task Resumption Using Visual Feedback". In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work*. ACM, 2014, pp. 767–777. DOI: 10.1145/2531602.2531710.

[181] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv* 1907.11692 (2019), pp. 1–13. DOI: 10.48550/arXiv.1907.11692.

[182] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. "Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 1–13. DOI: 10.1145/3313831.3376739.

[183] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. "Deep Photo Style Transfer". In: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 6997–7005. DOI: 10.1109/CVPR.2017.740.

[184] Saturnino Luz, Masood Masoodian, and Bill Rogers. "Supporting Collaborative Transcription of Recorded Speech with a 3D Game Interface". In: *Proceedings of the 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Vol. 6279. Springer Berlin Heidelberg, 2010, pp. 394–401. DOI: 10.1007/978-3-642-15384-6_42.

[185] Saturnino Luz, Masood Masoodian, Bill Rogers, and Chris Deering. "Interface Design Strategies for Computer-Assisted Speech Transcription". In: *Proceedings of the 20th Australasian Computer-Human Interaction Conference*. Vol. 287. ACM, 2008, pp. 203–210. DOI: 10.1145/1517744.1517812.

[186] Collin Lynch, Kevin D. Ashley, Niels Pinkwart, and Vincent Aleven. "Concepts, Structures, and Goals: Redefining Ill-Definedness". In: *International Journal of Artificial Intelligence in Education* 19.3 (2009), pp. 253–266.

[187] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data Using T-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.

[188] Norman H. Mackworth. "The Breakdown of Vigilance During Prolonged Visual Search". In: *Quarterly Journal of Experimental Psychology* 1.1 (1948), pp. 6–21. DOI: 10.1080/17470214808416738.

[189] Chandra Shekhar Maddila, Sai Surya Upadrasta, Chetan Bansal, Nachiappan Nagappan, Georgios Gousios, and Arie van Deursen. "Nudge: Accelerating Overdue Pull Requests Towards Completion". In: *ACM Transactions on Software Engineering and Methodology* 32.2 (2022), pp. 1–30. DOI: 10.1145/3544791.

[190] Eleanor A. Maguire, David G. Gadian, Ingrid S. Johnsrude, Catriona D. Good, John Ashburner, Richard S. J. Frackowiak, and Christopher D. Frith. "Navigation-Related Structural Change in the Hippocampi of Taxi Drivers". In: *Proceedings of the National Academy of Sciences* 97.8 (2000), pp. 4398–4403. DOI: 10.1073/pnas.070039597.

[191] George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. "Human-Centered Design of Artificial Intelligence". In: *Handbook of Human Factors and Ergonomics*. Wiley, 2021, pp. 1085–1106. DOI: 10.1002/9781119636113.ch42.

[192] Gloria Mark, Daniela Gudith, and Ulrich Klocke. "The Cost of Interrupted Work: More Speed and Stress". In: *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 107–110. DOI: 10.1145/1357054.1357072.

[193] Joe Marks, Brad Andalman, Paul A. Beardsley, William T. Freeman, Sarah F. Gibson, Jessica K. Hodgins, T. Kang, Brian Mirtich, Hanspeter Pfister, Wheeler Ruml, Kathy Ryall, Joshua E. Seims, and Stuart M. Shieber. "Design Galleries: a General Approach to Setting Parameters for Computer Graphics and Animation". In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 1997, pp. 389–400. DOI: 10.1145/258734.258887.

[194] Alison Marsh. "Respeaking for the BBC". In: *inTRAlinea* 1700 (2006).

[195] William D. Marslen-Wilson. "Speech Shadowing and Speech Comprehension". In: *Speech Communication* 4.1-3 (1985), pp. 55–73. DOI: 10.1016/0167-6393(85)90036-6.

[196] Fred G. Martin. "Will Massive Open Online Courses Change How We Teach?" In: *Communications of the ACM* 55.8 (2012), pp. 26–28. DOI: 10.1145/2240236.2240246.

[197] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and C. Raymond Perrault. "Artificial Intelligence Index Report 2023". In: *arXiv* 2310.03715 (2023), pp. 1–386. DOI: 10.48550/arxiv.2310.03715.

[198] Paul M Mastrangelo, Wendi Everton, and Jeffery A Jolton. "Personal Use of Work Computers: Distraction Versus Destruction". In: *CyberPsychology & Behavior* 9.6 (2006), pp. 730–741.

[199] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys* 54.115 (2022), pp. 1–35. DOI: 10.1145/3457607.

[200] Lise Menn and Nan Bernstein Ratner, eds. *Methods for Studying Language Production*. Hove, UK: Psychology Press, 1999. DOI: 10.4324/9781410601599.

[201] Juan Daniel Valor Miró, Joan Albert Silvestre-Cerdà, Jorge Civera, Carlos Turró, and Alfons Juan. "Efficiency and Usability Study of Innovative Computer-Aided Transcription Strategies for Video Lecture Repositories". In: *Speech Communication* 74 (2015), pp. 65–75. DOI: 10.1016/j.specom.2015.09.006.

[202] Gautam Mittal, Jesse H. Engel, Curtis Hawthorne, and Ian Simon. "Symbolic Music Generation with Diffusion Models". In: *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021, pp. 468–475.

[203] Diego Montes, Pongpatapee Peerapatanapokin, Jeff Schultz, Chengjun Guo, Wenxin Jiang, and James C. Davis. "Discrepancies Among Pre-Trained Deep Neural Networks: A New Threat To Model Zoo Reliability". In: *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, 2022, pp. 1605–1609. DOI: 10.1145/3540250.3560881.

[204] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, and José Bobes-Bascarán. "A Classification and Review of Tools for Developing and Interacting with Machine Learning Systems". In: *Proceedings of the 37th ACM SIGAPP Symposium on Applied Computing*. ACM, 2022, pp. 1092–1101. DOI: 10.1145/3477314.3507310.

[205] Jack Mostow and Gregory Aist. "Evaluating Tutors That Listen: An Overview of Project LISTEN". In: *Smart Machines in Education: The Coming Revolution in Educational Technology*. Cambridge, MA: MIT Press, 2001, pp. 169–234.

[206] Michael J. Muller, Lydia B. Chilton, Anna Kantosalo, Charles Patrick Martin, and Greg Walsh. "Proceedings of the GenAICHI Workshop: Generative AI and HCI". In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. 110. ACM, 2022, pp. 1–7. DOI: 10.1145/3491101.3503719.

[207] Nadia Nahar, Shurui Zhou, Grace A. Lewis, and Christian Kästner. "Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process". In: *Proceedings of the 44th IEEE/ACM International Conference on Software Engineering*. ACM, 2022, pp. 413–425. DOI: 10.1145/3510003.3510209.

[208] Shin'ya Nakajima and James F. Allen. "A Study on Prosody and Discourse Structure in Cooperative Dialogues". In: *Phonetica* 50.3 (1993), pp. 197–210. DOI: 10.1159/000261940.

[209] Ali Bou Nassif, Ismail Shahin, Imtinan Basem Attili, Mohammad Azzeh, and Khaled Shaalan. "Speech Recognition Using Deep Neural Networks: A Systematic Review". In: *IEEE Access* 7 (2019), pp. 19143–19165. DOI: 10.1109/ACCESS.2019.2896880.

[210] Peter Nemenyi. "Distribution-Free Multiple Comparisons". PhD thesis. Princeton University, 1963.

[211] Stephen C. Nettelhorst and Laura A. Brannon. "The Effect of Advertisement Choice, Sex, and Need for Cognition on Attention". In: *Computers in Human Behavior* 28.4 (2012), pp. 1315–1320. DOI: 10.1016/j.chb.2012.02.015.

[212] Pedro Neves, José Fornari, and João Batista Florindo. "Generating Music with Sentiment Using Transformer-GANs". In: *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. ISMIR, 2022, pp. 717–725.

[213] Kee Yuan Ngiam and Ing Wei Khor. "Big Data and Machine Learning Algorithms for Health-Care Delivery". In: *The Lancet Oncology* 20.5 (2019), e262–e273. DOI: 10.1016/s1470-2045(19)30149-4.

[214] Giang Nguyen, Stefan Dlugolinsky, Martin Bobák, Viet D. Tran, Álvaro López García, Ignacio Heredia, Peter Malík, and Ladislav Hluchý. "Machine Learning and Deep Learning Frameworks and Libraries for Large-Scale Data Mining: A Survey". In: *Artificial Intelligence Review* 52.1 (2019), pp. 77–124. DOI: 10.1007/S10462-018-09679-Z.

[215] Donald A. Norman. *Design for a Better World: Meaningful, Sustainable, Humanity Centered*. Cambridge, US: MIT Press, 2023.

[216] Donald A. Norman. "Human-Centered Design Considered Harmful". In: *Interactions* 12.4 (2005), pp. 14–19. DOI: 10.1145/1070960.1070976.

[217] Donald A. Norman and Stephen W. Draper. *User Centered System Design; New Perspectives on Human-Computer Interaction*. Mahwah, US: Lawrence Erlbaum Associates, 1986.

[218] Michael O'Neill and Mark Connor. "Amplifying Limitations, Harms and Risks of Large Language Models". In: *arXiv* 2307.04821 (2023), pp. 1–19. DOI: 10.48550/arxiv.2307.04821.

[219] Rachel S. Oeppen, Graham Shaw, and Peter A. Brennan. "Human Factors Recognition at Virtual Meetings and Video Conferencing: How to Get the Best Performance From Yourself and Others". In: *British Journal of Oral and Maxillofacial Surgery* 58.6 (2020), pp. 643–646. DOI: 10.1016/j.bjoms.2020.04.046.

[220] Jun Ogata, Masataka Goto, and Kouichirou Eto. "Automatic Transcription for a Web 2.0 Service to Search Podcasts". In: *Proceedings of the 8th Annual Conference of the International Speech Communication Association*. ISCA, 2007, pp. 2617–2620.

[221] Mayu Omiya, Yusuke Horiuchi, Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. "Optimization-Based Data Generation for Photo Enhancement". In: *Proceedings of the 4th IEEE Workshop on New Trends in Image Restoration and Enhancement*. IEEE, 2019, pp. 1890–1898. DOI: 10.1109/CVPRW.2019.00240.

[222] OpenAI. "GPT-4 Technical Report". In: *arXiv* 2303.08774 (2023), pp. 1–100. DOI: 10.48550/arxiv.2303.08774.

[223] Jonas Oppenlaender. "A Taxonomy of Prompt Modifiers for Text-to-Image Generation". In: *Behaviour & Information Technology* (2023), pp. 1–14. DOI: 10.1080/0144929x.2023.2286532.

[224] Jonas Oppenlaender, Rhema Linder, and Johanna M. Silvennoinen. "Prompting AI Art: An Investigation Into the Creative Skill of Prompt Engineering". In: *arXiv* 2303.13534 (2023), pp. 1–34. DOI: 10.48550/arXiv.2303.13534.

[225] Raja Parasuraman. "Memory Load and Event Rate Control Sensitivity Decrements in Sustained Attention". In: *Science* 205.4409 (1979), pp. 924–927.

[226] Marco Pasini and Jan Schlüter. "Musika! Fast Infinite Waveform Music Generation". In: *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. ISMIR, 2022, pp. 543–550.

[227] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. "Data and Its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research". In: *Patterns* 2.11 (2021), p. 100336. DOI: 10.1016/J.PATTER.2021.100336.

[228] Evandro Morais Peixoto, Ana Celi Pallini, Robert J Vallerand, Sonia Rahimi, and Marcus Vinicius Silva. "The Role of Passion for Studies on Academic Procrastination and Mental Health During the COVID-19 Pandemic". In: *Social Psychology of education* 24.3 (2021), pp. 877–893. DOI: 10.1007/s11218-021-09636-9.

[229] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. "Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture". In: *Proceedings of the 2018 IEEE Spoken Language Technology Workshop*. IEEE, 2018, pp. 513–520. DOI: 10.1109/SLT.2018.8639643.

[230] Richard E. Petty and John T. Cacioppo. "The Elaboration Likelihood Model of Persuasion". In: *Advances in Experimental Social Psychology*. Elsevier, 1986, pp. 123–205. DOI: 10.1016/s0065-2601(08)60214-2.

[231] Jason Phang, Thibault Févry, and Samuel R. Bowman. "Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-Data Tasks". In: *arXiv* 1811.01088 (2018), pp. 1–12. DOI: 10.48550/arXiv.1811.01088.

[232] Robert Philipp, Andreas Mladenow, Christine Strauss, and Alexander Völz. "Machine Learning as a Service: Challenges in Research and Applications". In: *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services*. ACM, 2020, pp. 396–406. DOI: 10.1145/3428757.3429152.

[233] François Pitié, Anil C. Kokaram, and Rozenn Dahyot. "Automated Colour Grading Using Colour Distribution Transfer". In: *Computer Vision and Image Understanding* 107.1-2 (2007), pp. 123–137. DOI: 10.1016/J.CVIU.2006.11.011.

[234] Peter G. Polson and Clayton H. Lewis. "Theory-Based Design for Easily Learned Interfaces". In: *Human-Computer Interaction* 5.2-3 (1990), pp. 191–220. DOI: 10.1080/07370024.1990.9667154.

[235] Fernando Poyatos. *Paralanguage: A Linguistic and Interdisciplinary Approach to Interactive Speech and Sounds*. Amsterdam, Netherlands: John Benjamins Publishing Company, 1993. ISBN: 9-027-23527-9.

[236] Aleš Pražák, Zdeněk Loose, Josef V. Psutka, Vlasta Radová, and Josef Psutka. "Live TV Subtitling Through Respeaking with Remote Cutting-Edge Technology". In: *Multimedia Tools and Applications* 79.1–2 (2020), pp. 1203–1220. DOI: 10.1007/s11042-019-08235-3.

[237] Alberto Prieto, Beatriz Prieto, Eva M. Ortigosa, Eduardo Ros, Francisco J. Pelayo, Julio Ortega, and Ignacio Rojas. "Neural Networks: An Overview of Early Research, Current Frameworks and New Challenges". In: *Neurocomputing* 214 (2016), pp. 242–268. DOI: 10.1016/J.NEUCOM.2016.06.014.

[238] Vinay Pursnani, Yusuf Sermet, Musa Kurt, and Ibrahim Demir. "Performance of ChatGPT on the US Fundamentals of Engineering Exam: Comprehensive Assessment of Proficiency and Potential Implications for Professional Environmental Engineering Practice". In: *Computers and Education: Artificial Intelligence* 5 (2023), p. 100183. DOI: 10.1016/j.caeai.2023.100183.

[239] Timothy A Pychyl. *Solving the Procrastination Puzzle: A Concise Guide to Strategies for Change*. New York, NY: TarcherPerigee, 2013.

[240] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67.

[241] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. "Machine Behaviour". In: *Nature* 568.7753 (2019), pp. 477–486. DOI: 10.1038/s41586-019-1138-y.

[242] Simeon Rau, Frank Heyen, Stefan Wagner, and Michael Sedlmair. "Visualization for AI-Assisted Composing". In: *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. ISMIR, 2022, pp. 151–159. DOI: 10.5281/zenodo.7316618.

[243] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. "Color Transfer Between Images". In: *IEEE Computer Graphics and Applications* 21.5 (2001), pp. 34–41. DOI: 10.1109/38.946629.

[244] Laria Reynolds and Kyle McDonell. "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 314. ACM, 2021, pp. 1–7. DOI: 10.1145/3411763.3451760.

[245] Mauro Ribeiro, Katarina Grolinger, and Miriam A. M. Capretz. "MLaaS: Machine Learning as a Service". In: *Proceedings of the 14th IEEE International Conference on Machine Learning and Applications*. IEEE, 2015, pp. 896–902. DOI: 10.1109/ICMLA.2015.152.

[246] Curtis Roads. "Research in Music and Artificial Intelligence". In: *ACM Computing Surveys* 17.2 (1985), pp. 163–190. DOI: 10.1145/4468.4469.

[247] Joaquin Alfredo Rodriguez, Gabriele Piccoli, and Marcin Bartosiak. "Nudging the Classroom: Designing a Socio-Technical Artifact to Reduce Academic Procrastination". In: *Proceedings of the 52nd Hawaii International Conference on System Sciences*. ScholarSpace, 2019, pp. 1–10.

[248] Melissa Roemmele and Andrew S. Gordon. "Creative Help: A Story Writing Assistant". In: *Proceedings of the 8th International Conference on Interactive Digital Storytelling*. Vol. 9445. Springer, 2015, pp. 81–92. DOI: 10.1007/978-3-319-27036-4_8.

[249] Martin Rohrmeier. "On Creativity, Music's AI Completeness, and Four Challenges for Artificial Musical Creativity". In: *Transactions of the International Society for Music Information Retrieval* 5.1 (2022), pp. 50–66. DOI: 10.5334/tismir.104.

[250] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 10674–10685. DOI: 10.1109/CVPR52688.2022.01042.

[251] Pablo Romero-Fresco. "Respeaking in Translator Training Curricula". In: *The Interpreter and Translator Trainer* 6.1 (2012), pp. 91–112. DOI: 10.1080/13556509.2012.10798831.

[252] Pablo Romero-Fresco. *Subtitling Through Speech Recognition: Respeaking*. London, UK: Routledge, 2020. DOI: 10.4324/9781003073147.

[253] Pablo Romero-Fresco and Carlo Eugeni. "Live Subtitling Through Respeaking". In: *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*. Springer International Publishing, 2020, pp. 269–295. DOI: 10.1007/978-3-030-42105-2_14.

[254] Drew S. Roselli, Jeanna N. Matthews, and Nisha Talagala. "Managing Bias in AI". In: *Companion Proceedings of the 2019 World Wide Web Conference*. ACM, 2019, pp. 539–544. DOI: 10.1145/3308560.3317590.

[255] Steven I. Ross, Fernando Martinez, Stephanie Houde, Michael J. Muller, and Justin D. Weisz. "The Programmer's Assistant: Conversational Interaction with a Large Language Model for Software Development". In: *Proceedings of the 28th ACM International Conference on Intelligent User Interfaces*. ACM, 2023, pp. 491–514. DOI: 10.1145/3581641.3584037.

[256] James A. Russell. "A Circumplex Model of Affect". In: *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1161–1178. DOI: 10.1037/h0077714.

[257] C. Douglas Saddler and Laurie A. Sacks. "Multidimensional Perfectionism and Academic Procrastination: Relationships with Depression in University Student". In: *Psychological Reports* 73.3 (1 1993), pp. 863–871. DOI: 10.1177/00332941930733pt123.

[258] Julie Saint-Lot, Jean-Paul Imbert, and Frédéric Dehais. "Red Alert: A Cognitive Countermeasure to Mitigate Attentional Tunneling". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 580. ACM, 2020, pp. 1–6. DOI: 10.1145/3313831.3376709.

[259] Kathleen C. Salter and Richard F. Fawcett. "The ART Test of Interaction: A Robust and Powerful Rank Test of Interaction in Factorial Models". In: *Communications in Statistics – Simulation and Computation* 22.1 (1993), pp. 137–153. DOI: 10.1080/03610919308813085.

[260] Timo Schick, Jane A. Yu, Zhengbao Jiang, Fabio Petroni, Patrick S. H. Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. "PEER: A Collaborative Language Model". In: *Proceedings of the 11th International Conference on Learning Representations*. OpenReview.net, 2023.

[261] Henri C. Schouwenburg. "Procrastinators and Fear of Failure: An Exploration of Reasons for Procrastination". In: *European Journal of Personality* 6.3 (1992), pp. 225–236. DOI: 10.1002/per.2410060305.

[262] Michael Sedlmair, Christoph Heinzl, Stefan Bruckner, Harald Piringer, and Torsten Möller. "Visual Parameter Space Analysis: A Conceptual Framework". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2161–2170. DOI: 10.1109/TVCG.2014.2346321.

[263] Athar Sefid, Jian Wu, Prasenjit Mitra, and C. Lee Giles. "Automatic Slide Generation for Scientific Papers". In: *Proceedings of the 3rd International Workshop on Capturing Scientific Knowledge*. Vol. 2526. CEUR-WS.org, 2019, pp. 11–16.

[264] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. "Taking the Human Out of the Loop: A Review of Bayesian Optimization". In: *Proceedings of the IEEE* 104.1 (2016), pp. 148–175. DOI: 10.1109/JPROC.2015.2494218.

[265] Hanieh Shakeri, Carman Neustaedter, and Steve DiPaola. "SAGA: Collaborative Storytelling with GPT-3". In: *Companion Publication of the 2021 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2021, pp. 163–166. DOI: 10.1145/3462204.3481771.

[266] Kshitij Sharma, Hamed S. Alavi, Patrick Jermann, and Pierre Dillenbourg. "A Gaze-Based Learning Analytics Model: in-Video Visual Feedback to Improve Learner'S Attention in MOOCs". In: *Proceedings of the 6th ACM International Conference on Learning Analytics & Knowledge*. ACM, 2016, pp. 417–421. DOI: 10.1145/2883851.2883902.

[267] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. "The Woman Worked as a Babysitter: On Biases in Language Generation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. ACL, 2019, pp. 3405–3410. DOI: 10.18653/v1/D19-1339.

[268] Chet C. Sherwood and Aida Gómez-Robles. "Brain Plasticity and Human Evolution". In: *Annual Review of Anthropology* 46.1 (2017), pp. 399–419. DOI: 10.1146/annurev-anthro-102215-100009.

[269] Liang Shi, Beichen Li, Milos Hasan, Kalyan Sunkavalli, Tamy Boubekeur, Radomír Mech, and Wojciech Matusik. "MATch: Differentiable Material Graphs for Procedural Material Capture". In: *ACM Transactions on Graphics* 39.196 (2020), pp. 1–15. DOI: 10.1145/3414685.3417781.

[270] Ben Shneiderman. "Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems". In: *ACM Transactions on Interactive Intelligent Systems* 10.26 (2020), pp. 1–31. DOI: 10.1145/3419764.

[271] Ben Shneiderman. *Human-Centered AI*. Oxford, UK: Oxford University Press, 2022.

[272] Anton Sigfrids, Jaana Leikas, Henrikki Salo-Pöntinen, and Emmi Koskimies. "Human-Centricity in AI Governance: A Systemic Approach". In: *Frontiers in Artificial Intelligence* 6.976887 (2023), pp. 1–9. DOI: 10.3389/frai.2023.976887.

[273] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. "Mastering the Game of Go with Deep Neural Networks and Tree Search". In: *Nature* 529.7587 (2016), pp. 484–489. DOI: 10.1038/nature16961.

[274] Ian Simon, Dan Morris, and Sumit Basu. "MySong: Automatic Accompaniment Generation for Vocal Melodies". In: *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 725–734. DOI: 10.1145/1357054.1357169.

[275] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. "Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence". In: *ACM Transactions on Computer-Human Interaction* 30.5 (2023), pp. 1–57. DOI: 10.1145/3511599.

[276] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 132–157. DOI: 10.1109/TASLP.2020.3038524.

[277] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. "A Statistical Model-Based Voice Activity Detection". In: *IEEE Signal Processing Letter* 6.1 (1999), pp. 1–3. DOI: 10.1109/97.736233.

[278] Konrad Sowa, Aleksandra Przegalinska, and Leon Ciechanowski. "Cobots in Knowledge Work". In: *Journal of Business Research* 125 (2021), pp. 135–142. DOI: 10.1016/j.jbusres.2020.11.038.

[279] Matthias Sperber, Graham Neubig, Christian Fügen, Satoshi Nakamura, and Alex Waibel. "Efficient Speech Transcription Through Respeaking". In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association*. ISCA, 2013, pp. 1087–1091.

[280] Matthias Sperber, Graham Neubig, Satoshi Nakamura, and Alex Waibel. "Optimizing Computer-Assisted Transcription Quality with Iterative User Interfaces". In: *Proceedings of the 10th International Conference on Language Resources and Evaluation*. ELRA, 2016, pp. 1986–1992.

[281] Constantin Spille, Birger Kollmeier, and Bernd T. Meyer. "Comparing Human and Automatic Speech Recognition in Simple and Complex Acoustic Scenes". In: *Computer Speech and Language* 52 (2018), pp. 123–140. DOI: 10.1016/j.csl.2018.04.003.

[282] Erik Stolterman and Mikael Wiberg. "Concept-Driven Interaction Design Research". In: *Human-Computer Interaction* 25.2 (2010), pp. 95–118. DOI: 10.1080/07370020903586696.

[283] Cary Stothart, Ainsley Mitchum, and Courtney Yehnert. "The Attentional Cost of Receiving a Cell Phone Notification". In: *Journal of Experimental Psychology: Human Perception and Performance* 41.4 (2015), pp. 893–897. DOI: 10.1037/xhp0000100.

[284] Anselm Strauss and Juliet Corbin. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Newbury Park, CA: Sage Publications, 1990.

[285] Jan Willem Streefkerk, Myra P. van Esch-Bussemakers, and Mark A. Neerincx. "Field Evaluation of a Mobile Location-Based Notification System for Police Officers". In: *Proceedings of the 10th Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 2008, pp. 101–108. DOI: 10.1145/1409240.1409252.

[286] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics*. ACL, 2019, pp. 3645–3650. DOI: 10.18653/V1/P19-1355.

[287] Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, and Carrie J. Cai. "AI as Social Glue: Uncovering the Roles of Deep Generative AI During Social Music Composition". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 582. ACM, 2021, pp. 1–11. DOI: 10.1145/3411764.3445219.

[288] Joseph W. Sullivan and Frances Degen Horowitz. "The Effects of Intonation on Infant Attention: the Role of the Rising Intonation Contour". In: *Journal of Child Language* 10.3 (1983), pp. 521–534. DOI: 10.1017/s0305000900005341.

[289] Harini Suresh and John V. Guttag. "A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle". In: *Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 17. ACM, 2021, pp. 1–9. DOI: 10.1145/3465416.3483305.

[290] Keith Swanwick. *Musical Knowledge: Intuition, Analysis and Music Education*. London, UK: Routledge, 2002. DOI: 10.4324/9780203424575.

[291] Marc Swerts. "Prosodic Features at Discourse Boundaries of Different Strength". In: *The Journal of the Acoustical Society of America* 101.1 (1997), pp. 514–521. DOI: 10.1121/1.418114.

[292] Hideyuki Takagi. "Interactive Evolutionary Computation: Fusion of the Capabilities of EC Optimization and Human Evaluation". In: *Proceedings of the IEEE* 89.9 (2001), pp. 1275–1296. DOI: 10.1109/5.949485.

[293] Jerry O. Talton, Daniel Gibson, Lingfeng Yang, Pat Hanrahan, and Vladlen Koltun. "Exploratory Modeling with Collaborative Design Spaces". In: *ACM Transactions on Graphics* 28.5 (2009), p. 167. DOI: 10.1145/1618452.1618513.

[294] Jerry O. Talton, Yu Lou, Steve Lesser, Jared Duke, Radomír Mech, and Vladlen Koltun. "Metropolis Procedural Modeling". In: *ACM Transactions on Graphics* 30.11 (2011), pp. 1–14. DOI: 10.1145/1944846.1944851.

[295] Hao Hao Tan and Dorien Herremans. "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling". In: *Proceedings of the 21st International Society for Music Information Retrieval Conference*. ISMIR, 2020, pp. 109–116.

[296] Michael A. Terry and Elizabeth D. Mynatt. "Side Views: Persistent, on-Demand Previews for Open-Ended Tasks". In: *Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2002, pp. 71–80. DOI: 10.1145/571985.571996.

[297] Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven, CT: Yale University Press, 2008.

[298] Chinchu Thomas and Dinesh Babu Jayagopi. "Predicting Student Engagement in Classrooms Using Facial Behavioral Cues". In: *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*. ACM, 2017, pp. 33–40. DOI: 10.1145/3139513.3139514.

[299] Tomoki Toda, Takashi Muramatsu, and Hideki Banno. "Implementation of Computationally Efficient Real-Time Voice Conversion". In: *Proceedings of the 13th Annual Conference of the International Speech Communication Association*. ISCA, 2012, pp. 94–97.

[300] Michael Tomasello. *The Cultural Origins of Human Cognition*. Cambridge, US: Harvard University Press, 1999. DOI: 10.2307/j.ctvjsf4jc.

[301] George L. Trager. "Paralanguage: A First Approximation". In: *Studies in Linguistics* 13 (1958), pp. 1–12.

[302] Viet Anh Trinh and Michael I. Mandel. "Directly Comparing the Listening Strategies of Humans and Machines". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 312–323. DOI: 10.1109/TASLP.2020.3040545.

[303] Emiru Tsunoo, Kentaro Shibata, Chaitanya Narisetty, Yosuke Kashiwagi, and Shinji Watanabe. "Data Augmentation Methods for End-to-End Speech Recognition on Distant-Talk Scenarios". In: *Proceedings of the 22nd Annual Conference of the International Speech Communication Association*. ISCA, 2021, pp. 301–305. DOI: 10.21437/INTERSPEECH.2021-958.

[304] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. "How It Works: A Field Study of Non-Technical Users Interacting with An Intelligent System". In: *Proceedings of the 2007 CHI Conference on Human Factors in Computing Systems*. ACM, 2007, pp. 31–40. DOI: 10.1145/1240624.1240630.

[305] Lisa Tweedie. "Interactive Visualisation Artifacts: How Can Abstractions Inform Design?" In: *Proceedings of the 10th BCS Conference on Human-Computer Interaction*. Cambridge University Press, 1995, pp. 247–265.

[306] Aditya Vashistha, Abhinav Garg, and Richard J. Anderson. "ReCall: Crowdsourcing on Basic Phones to Financially Sustain Voice Forums". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 169. ACM, 2019, pp. 1–13. DOI: 10.1145/3290605.3300399.

[307] Aditya Vashistha, Pooja Sethi, and Richard J. Anderson. "BSpeak: An Accessible Voice-Based Crowdsourcing Marketplace for Low-Income Blind People". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 57. ACM, 2018, pp. 1–13. DOI: 10.1145/3173574.3173631.

[308] Aditya Vashistha, Pooja Sethi, and Richard J. Anderson. "Respeak: A Voice-Based, Crowd-Powered Speech Transcription System". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 1855–1866. DOI: 10.1145/3025453.3025640.

[309] Narayanan Veliyath, Pradipta De, Andrew A. Allen, Charles B. Hodges, and Aniruddha Mitra. "Modeling Students' Attention in the Classroom Using Eyetrackers". In: *Proceedings of the 2019 ACM Southeast Regional Conference*. ACM, 2019, pp. 2–9. DOI: 10.1145/3299815.3314424.

[310] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. "User Acceptance of Information Technology: Toward a Unified View". In: *MIS Quarterly* 27.3 (2003), pp. 425–478. DOI: 10.2307/30036540.

[311] Roelof Anne Jelle de Vries, Cristina Zaga, Franciszka Bayer, Constance H. C. Drossaert, Khiet P. Truong, and Vanessa Evers. "Experts Get Me Started, Peers Keep Me Going: Comparing Crowd- Versus Expert-Designed Motivational Text Messages for Exercise Behavior Change". In: *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM, 2017, pp. 155–162. DOI: 10.1145/3154862.3154875.

[312] Luuk Waes, Mariëlle Leijten, and Aline Remael. "Live Subtitling with Speech Recognition: Causes and Consequences of Text Reduction". In: *Across Languages and Cultures* 14.1 (2013), pp. 15–46. DOI: 10.1556/acr.14.2013.1.2.

[313] Robert A. Wagner and Michael J. Fischer. "The String-to-String Correction Problem". In: *Journal of the ACM* 21.1 (1974), pp. 168–173. DOI: 10.1145/321796.321811.

[314] Bin Wang, Yukun Liu, Jing Qian, and Sharon K. Parker. "Achieving Effective Remote Working During the COVID-19 Pandemic: A Work Design Perspective". In: *Applied Psychology* 70.1 (2020), pp. 16–59. DOI: 10.1111/apps.12290.

[315] Kai Wang and Jeffrey V. Nickerson. "A Literature Review on Individual Creativity Support Systems". In: *Computers in Human Behavior* 74 (2017), pp. 139–151. DOI: 10.1016/j.chb.2017.04.035.

[316] Xiangdong Wang, Ying Yang, Hong Liu, and Yueliang Qian. "Improving Speech Transcription by Exploiting User Feedback and Word Repetition". In: *Multimedia Tools and Applications* 76.19 (2017), pp. 20359–20376. DOI: 10.1007/s11042-017-4714-x.

[317] Ziyu Wang, Dingsu Wang, Yixiao Zhang, and Gus Xia. "Learning Interpretable Representation for Controllable Polyphonic Music Generation". In: *Proceedings of the 21st International Society for Music Information Retrieval Conference*. ISMIR, 2020, pp. 662–669.

[318] Ziyu Wang and Gus Xia. "MuseBERT: Pre-Training Music Representation for Music Understanding and Controllable Generation". In: *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021, pp. 722–729.

[319] Chatchai Wangwiwattana, Sunjoli Aggarwal, and Eric C. Larson. "Writers Gonna Wait: The Effectiveness of Notifications to Initiate Aversive Action in Writing Procrastination". In: *arXiv* 2101.10191 (2021), pp. 1–13. DOI: 10.48550/arXiv.2101.10191.

[320] Brian Wansink and Koert van Ittersum. "Portion Size Me: Downsizing Our Consumption Norms". In: *Journal of the American Dietetic Association* 107.7 (2007), pp. 1103–1106. DOI: 10.1016/j.jada.2007.05.019.

[321] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin andJahn Heymann andMatthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. "ESPnet: End-to-End Speech Processing Toolkit". In: *Proceedings of the 19th Annual Conference of the International Speech Communication Association*. ISCA, 2018, pp. 2207–2211. DOI: 10.21437/Interspeech.2018-1456.

[322] Shinji Watanabe, Michael I. Mandel, Jon Barker, and Emmanuel Vincent. "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings". In: *arXiv* 2004.09249 (2020), pp. 1–7. DOI: 10.48550/arXiv.2004.09249.

[323] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. "Emergent Abilities of Large Language Models". In: *Transactions on Machine Learning Research* (2022).

[324] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. "Taxonomy of Risks Posed by Language Models". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2022, pp. 214–229. DOI: 10.1145/3531146.3533088.

[325] Joyce Weiner. *Why AI/Data Science Projects Fail: How to Avoid Project Pitfalls*. Cham, Switzerland: Springer International Publishing, 2021. DOI: 10.1007/978-3-031-01685-1.

[326] Mark Weiser and John Seely Brown. *Designing Calm Technology*. Tech. rep. Palo Alto, CA: Xerox PARC, 1995, pp. 1–5.

[327] Mark Weiser and John Seely Brown. "The Coming Age of Calm Technology". In: *Beyond calculation*. Springer, 1997, pp. 75–85. DOI: 10.1007/978-1-4612-0685-9_6.

[328] Steve Whittaker. "Rethinking Video as a Technology for Interpersonal Communications: Theory and Design Implications". In: *International Journal of Human-Computer Studies* 42.5 (1995), pp. 501–529. DOI: 10.1006/ijhc.1995.1022.

[329] Steve Whittaker, Loren G. Terveen, and Bonnie A. Nardi. "Let's Stop Pushing the Envelope and Start Addressing It: A Reference Task Agenda for HCI". In: *Human-Computer Interaction* 15.2–3 (2000), pp. 75–106. DOI: 10.1207/S15327051HCI1523_2.

[330] Christopher D. Wickens. "Attention: Theory, Principles, Models and Applications". In: *International Journal of Human-Computer Interaction* 37.5 (2021), pp. 403–417. DOI: 10.1080/10447318.2021.1874741.

[331] Austin P. Wright, Zijie J. Wang, Haekyu Park, Grace Guo, Fabian Sperrle, Mennatallah El-Assady, Alex Endert, Daniel A. Keim, and Duen Horng Chau. "A Comparative Analysis of Industry Human-AI Interaction Guidelines". In: *Proceedings of the 2020 IEEE VIS Workshop on TRust and EXpertise in Visual Analytics*. arXiv, 2020, pp. 1–8. DOI: 10.48550/arXiv.2010.11761.

[332] Peter C. Wright and Andrew F. Monk. "The Use of Think-Aloud Evaluation Methods in Design". In: *ACM SIGCHI Bulletin* 23.1 (1991), pp. 55–57. DOI: 10.1145/122672.122685.

[333] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim M. Hazelwood. "Sustainable AI: Environmental Implications, Challenges and Opportunities". In: *Proceedings of the 5th Conference on Machine Learning and Systems*. mlsys.org, 2022, pp. 795–813.

[334] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. "AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 385. ACM, 2022, pp. 1–22. DOI: 10.1145/3491102.3517582.

[335] Peter R. Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J. Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, Leilani Gilpin, Piyush Khandelwal, Varun Kompella, HaoChih Lin, Patrick MacAlpine, Declan Oller, Takuma Seno, Craig Sherstan, Michael D. Thomure, Houmehr Aghabozorgi, Leon Barrett, Rory Douglas, Dion Whitehead, Peter Dürr, Peter Stone, Michael Spranger, and Hiroaki Kitano. "Outracing Champion Gran Turismo Drivers with Deep Reinforcement Learning". In: *Nature* 602.7896 (2022), pp. 223–228. DOI: 10.1038/s41586-021-04357-7.

[336] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. "ELEGANT: Exchanging Latent Encodings with GAN for Transferring Multiple Face Attributes". In: *Proceedings of the 15th European Conference on Computer Vision*. Springer, 2018, pp. 172–187. DOI: 10.1007/978-3-030-01249-6_11.

[337] Xiang Xiao and Jingtao Wang. "Context and Cognitive State Triggered Interventions for Mobile MOOC Learning". In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 378–385. DOI: 10.1145/2993148.2993177.

[338] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L. Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. "Toward Human Parity in Conversational Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12 (2017), pp. 2410–2423. DOI: 10.1109/TASLP.2017.2756440.

[339] Wei Xu, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao. "Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI". In: *International Journal of Human-Computer Interaction* 39.3 (2023), pp. 494–518. DOI: 10.1080/10447318.2022.2041900.

[340] Yi Xu. "Speech Melody as Articulatorily Implemented Communicative Functions". In: *Speech Communication* 46.3-4 (2005), pp. 220–251. DOI: 10.1016/j.specom.2005.02.014.

[341] Hiromu Yakura. "No More Handshaking: How Have COVID-19 Pushed the Expansion of Computer-Mediated Communication in Japanese Idol Culture?" In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 645. ACM, 2021, pp. 1–10. DOI: 10.1145/3411764.3445252.

[342] Hiromu Yakura, Youhei Akimoto, and Jun Sakuma. "Generate (Non-Software) Bugs to Fool Classifiers". In: *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press, 2020, pp. 1070–1078. DOI: 10.1609/AAAI.V34I01.5457.

[343] Hiromu Yakura and Masataka Goto. "Enhancing Participation Experience in VR Live Concerts by Improving Motions of Virtual Audience Avatars". In: *Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2020, pp. 555–565. DOI: 10.1109/ISMAR50242.2020.00083.

[344] Hiromu Yakura and Masataka Goto. "IteraTTA: An Interface for Exploring Both Text Prompts and Audio Priors in Generating Music with Text-to-Audio Models". In: *Proceedings of the 24th International Society for Music Information Retrieval Conference*. ISMIR, 2023, pp. 129–137. DOI: 10.5281/ZENODO.10265239.

[345] Hiromu Yakura, Yuki Koyama, and Masataka Goto. "Tool- and Domain-Agnostic Parameterization of Style Transfer Effects Leveraging Pretrained Perceptual Metrics". In: *Proceedings of the 38th International Joint Conference on Artificial Intelligence*. ijcai.org, 2021, pp. 1208–1216. DOI: 10.24963/IJCAI.2021/167.

[346] Hiromu Yakura, Tomoyasu Nakano, and Masataka Goto. "An Automated System Recommending Background Music to Listen to While Working". In: *User Modeling and User-Adapted Interactions* 32.3 (2022), pp. 355–388. DOI: 10.1007/s11257-022-09325-y.

[347] Hiromu Yakura, Tomoyasu Nakano, and Masataka Goto. "FocusMusicRecommender: A System for Recommending Music to Listen to While Working". In: *Proceedings of the 23rd ACM International Conference on Intelligent User Interfaces*. ACM, 2018, pp. 7–17. DOI: 10.1145/3172944.3172981.

[348] Hiromu Yakura and Jun Sakuma. "Robust Audio Adversarial Example for a Physical Attack". In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. ijcai.org, 2019, pp. 5334–5341. DOI: 10.24963/IJCAI.2019/741.

[349] Hiromu Yakura, Shinnosuke Shinozaki, Reon Nishimura, Yoshihiro Oyama, and Jun Sakuma. "Malware Analysis of Imaged Binary Samples by Convolutional Neural Network with Attention Mechanism". In: *Proceedings of the 8th ACM Conference on Data and Application Security and Privacy*. ACM, 2018, pp. 127–134. DOI: 10.1145/3176258.3176335.

[350] Hiromu Yakura, Shinnosuke Shinozaki, Reon Nishimura, Yoshihiro Oyama, and Jun Sakuma. "Neural Malware Analysis with Attention Mechanism". In: *Computers and Security* 87.101592 (2019), pp. 1–15. DOI: 10.1016/J.COSE.2019.101592.

[351] Hiromu Yakura, Kento Watanabe, and Masataka Goto. "Self-Supervised Contrastive Learning for Singing Voices". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 1614–1623. DOI: 10.1109/TASLP.2022.3169627.

[352] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and Ray LC. "AI as an Active Writer: Interaction Strategies with Generated Text in Human-AI Collaborative Fiction Writing". In: *Proceedings of the 3rd ACM IUI Workshops on Human-AI Co-Creation with Generative Models*. CEUR-WS.org, 2022, pp. 56–65.

[353] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. "MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation". In: *Proceedings of the 18th International Society for Music Information Retrieval Conference*. ISMIR, 2017, pp. 324–331.

[354] Qian Yang, Nikola Banovic, and John Zimmerman. "Mapping Machine Learning Advances From HCI Research to Reveal Starting Places for Design Innovation". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 130. ACM, 2018, pp. 1–11. DOI: 10.1145/3173574.3173704.

[355] Qian Yang, Alex Sciuto, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. "Investigating How Experienced UX Designers Effectively Work with Machine Learning". In: *Proceedings of the 2018 ACM International Conference on Designing Interactive Systems*. ACM, 2018, pp. 585–596. DOI: 10.1145/3196709.3196730.

[356] Qian Yang, Aaron Steinfeld, Carolyn P. Rosé, and John Zimmerman. "Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 174. ACM, 2020, pp. 1–13. DOI: 10.1145/3313831.3376301.

[357] Jordan Yaniv, Yael Newman, and Ariel Shamir. "The Face of Art: Landmark Detection and Geometric Style in Portraits". In: *ACM Transactions on Graphics* 38.60 (2019), pp. 1–15. DOI: 10.1145/3306346.3322984.

[358] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda B. Viégas. "Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 356. ACM, 2023, pp. 1–13. DOI: 10.1145/3544548.3580900.

[359] Ming Yin, Jennifer Wortman Vaughan, and Hanna M. Wallach. "Understanding the Effect of Accuracy on Trust in Machine Learning Models". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. 279. DOI: 10.1145/3290605.3300509.

[360] Botao Yu, Peiling Lu, Rui Wang, Wei Hu, Xu Tan, Wei Ye, Shikun Zhang, Tao Qin, and Tie-Yan Liu. "Museformer: Transformer with Fine- and Coarse-Grained Attention for Music Generation". In: *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*. Curran Associates, 2022, pp. 1376–1388.

[361] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. "User Trust Dynamics: An Investigation Driven by Differences in System Performance". In: *Proceedings of the 22nd ACM International Conference on Intelligent User Interfaces*. ACM, 2017, pp. 307–317. DOI: 10.1145/3025171.3025219.

[362] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. "Wordcraft: Story Writing with Large Language Models". In: *Proceedings of the 27th ACM International Conference on Intelligent User Interfaces*. ACM, 2022, pp. 841–852. DOI: 10.1145/3490099.3511105.

[363] Mehmet Ersin Yümer and Niloy J. Mitra. "Spectral Style Transfer for Human Motion Between Independent Actions". In: *ACM Transactions on Graphics* 35.137 (2016), pp. 1–8. DOI: 10.1145/2897824.2925955.

[364] Janez Zaletelj and Andrej Kosir. "Predicting Students' Attention in the Classroom From Kinect Facial and Body Features". In: *EURASIP Journal on Image and Video Processing* 2017 (2017), p. 80. DOI: 10.1186/s13640-017-0228-8.

[365] Robert J. Zatorre and Jackson T. Gandour. "Neural Specializations for Speech and Pitch: Moving Beyond the Dichotomies". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1493 (2007), pp. 1087–1104. DOI: 10.1098/rstb.2007.2161.

[366] Sabah Zdanowska and Alex S. Taylor. "A Study of UX Practitioners Roles in Designing Real-World, Enterprise ML Systems". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 531. ACM, 2022, pp. 1–15. DOI: 10.1145/3491102.3517607.

[367] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. "StoryDrawer: A Child-AI Collaborative Drawing System to Support Children's Creative Visual Storytelling". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 311. ACM, 2022, pp. 1–15. DOI: 10.1145/3491102.3501914.

[368] Chao Zhang, Zili Zhou, Jiayi Wu, Yajing Hu, Yaping Shao, Jianhui Liu, Yuqi Hu, Fangtian Ying, and Cheng Yao. "Bio Sketchbook: An AI-Assisted Sketching Partner for Children's Biodiversity Observational Learning". In: *Proceedings of the 20th ACM Interaction Design and Children Conference*. ACM, 2021, pp. 466–470. DOI: 10.1145/3459990.3465197.

[369] Xinlei Zhang, Takashi Miyaki, and Jun Rekimoto. "withYou: Automated Adaptive Speech Tutoring with Context-Dependent Speech Recognition". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 195. ACM, 2020, pp. 1–12. DOI: 10.1145/3313831.3376322.

[370] Yixiao Zhang, Gus Xia, Mark Levy, and Simon Dixon. "COSMIC: A Conversational Interface for Human-AI Music Co-Creation". In: *Proceedings of the 21th International Conference on New Interfaces for Musical Expression*. nime.org, 2021. DOI: 10.21428/92fbeb44.110a7a32.

[371] Chengbo Zheng, Dakuo Wang, April Yi Wang, and Xiaojuan Ma. "Telling Stories From Computational Notebooks: AI-Assisted Presentation Slides Creation for Presenting Data Science Work". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 53. ACM, 2022, pp. 1–20. DOI: 10.1145/3491102.3517615.

[372] Jianlong Zhou, Syed Z. Arshad, Simon Luo, and Fang Chen. "Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making". In: *Proceedings of the 16th IFIP TC 13 International Conference on Human-Computer Interaction*. Springer, 2017, pp. 23–39. DOI: 10.1007/978-3-319-68059-0_2.

[373] Yijun Zhou, Yuki Koyama, Masataka Goto, and Takeo Igarashi. "Generative Melody Composition with Human-in-the-Loop Bayesian Optimization". In: *Proceedings of the 2020 Joint Conference on AI Music Creativity*. DiVA.org, 2020.

[374] Yijun Zhou, Yuki Koyama, Masataka Goto, and Takeo Igarashi. "Interactive Exploration-Exploitation Balancing for Generative Melody Composition". In: *Proceedings of the 26th ACM International Conference on Intelligent User Interfaces*. ACM, 2021, pp. 43–47. DOI: 10.1145/3397481.3450663.

[375] Albert Ziegler, Eirini Kalliamvakou, X. Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. "Productivity Assessment of Neural Code Completion". In: *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. ACM, 2022, pp. 21–29. DOI: 10.1145/3520312.3534864.

[376] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. "Research Through Design as a Method for Interaction Design Research in HCI". In: *Proceedings of the 2007 Conference on Human Factors in Computing Systems*. ACM, 2007, pp. 493–502. DOI: 10.1145/1240624.1240704.

[377] John Zimmerman, Erik Stolterman, and Jodi Forlizzi. "An Analysis and Critique of *Research Through Design*: Towards A Formalization of a Research Approach". In: *Proceedings of the 2010 ACM International Conference on Designing Interactive Systems*. ACM, 2010, pp. 310–319. DOI: 10.1145/1858171.1858228.

Publication list for the doctoral degree qualification

Peer-referred journal papers

[1] Hiromu Yakura, Tomoyasu Nakano, and Masataka Goto. "An Automated System Recommending Background Music to Listen to While Working". In: *User Modeling and User-Adapted Interactions* 32.3 (2022), pp. 355–388. DOI: 10.1007/s11257-022-09325-y.

[2] Hiromu Yakura, Kento Watanabe, and Masataka Goto. "Self-Supervised Contrastive Learning for Singing Voices". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 1614–1623. DOI: 10.1109/TASLP.2022.3169627.

[3] Hiromu Yakura, Shinnosuke Shinozaki, Reon Nishimura, Yoshihiro Oyama, and Jun Sakuma. "Neural Malware Analysis with Attention Mechanism". In: *Computers and Security* 87.101592 (2019), pp. 1–15. DOI: 10.1016/J.COSE.2019.101592.

Peer-reviewed conference papers

[1] Hiromu Yakura and Masataka Goto. "IteraTTA: An Interface for Exploring Both Text Prompts and Audio Priors in Generating Music with Text-to-Audio Models". In: *Proceedings of the 24th International Society for Music Information Retrieval Conference*. ISMIR, 2023, pp. 129–137. DOI: 10.5281/ZENODO.10265239.

[2] Riku Arakawa*, Hiromu Yakura*, and Masataka Goto. "CatAlyst: Domain-Extensible Intervention for Preventing Task Procrastination Using Large Generative Models". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 157. ACM, 2023, pp. 1–19. DOI: 10.1145/3544548.3581133.

[3] Riku Arakawa*, Hiromu Yakura*, and Masataka Goto. "BeParrot: Efficient Interface for Transcribing Unclear Speech via Respeaking". In: *Proceedings of the 27th ACM International Conference on Intelligent User Interfaces*. ACM, 2022, pp. 832–840. DOI: 10.1145/3490099.3511164.

[4] Hiromu Yakura, Yuki Koyama, and Masataka Goto. "Tool- and Domain-Agnostic Parameterization of Style Transfer Effects Leveraging Pretrained Perceptual Metrics". In: *Proceedings of the 38th International Joint Conference on Artificial Intelligence*. ijcai.org, 2021, pp. 1208–1216. DOI: 10.24963/IJCAI.2021/167.

[5] Riku Arakawa* and Hiromu Yakura*. "Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 99. ACM, 2021, pp. 1–15. DOI: 10.1145/3411764.3445339.

[6] Hiromu Yakura. "No More Handshaking: How Have COVID-19 Pushed the Expansion of Computer-Mediated Communication in Japanese Idol Culture?" In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 645. ACM, 2021, pp. 1–10. DOI: 10.1145/3411764.3445252.

*Equal contribution

[7] Hiromu Yakura and Masataka Goto. “Enhancing Participation Experience in VR Live Concerts by Improving Motions of Virtual Audience Avatars”. In: *Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2020, pp. 555–565. DOI: 10.1109/ISMAR50242.2020.00083.

[8] Riku Arakawa* and Hiromu Yakura*. “INWARD: A Computer-Supported Tool for Video-Reflection Improves Efficiency and Effectiveness in Executive Coaching”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 574. ACM, 2020, pp. 1–13. DOI: 10.1145/3313831.3376703.

[9] Hiromu Yakura, Youhei Akimoto, and Jun Sakuma. “Generate (Non-Software) Bugs to Fool Classifiers”. In: *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press, 2020, pp. 1070–1078. DOI: 10.1609/AAAI.V34I01.5457.

[10] Hiromu Yakura and Jun Sakuma. “Robust Audio Adversarial Example for a Physical Attack”. In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. ijcai.org, 2019, pp. 5334–5341. DOI: 10.24963/IJCAI.2019/741.

[11] Riku Arakawa* and Hiromu Yakura*. “REsCUE: A Framework for REal-Time Feedback on Behavioral CUEs Using Multimodal Anomaly Detection”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 572. ACM, 2019, pp. 1–13. DOI: 10.1145/3290605.3300802.

[12] Hiromu Yakura, Shinnosuke Shinozaki, Reon Nishimura, Yoshihiro Oyama, and Jun Sakuma. “Malware Analysis of Imaged Binary Samples by Convolutional Neural Network with Attention Mechanism”. In: *Proceedings of the 8th ACM Conference on Data and Application Security and Privacy*. ACM, 2018, pp. 127–134. DOI: 10.1145/3176258.3176335.

[13] Hiromu Yakura, Tomoyasu Nakano, and Masataka Goto. “FocusMusicRecommender: A System for Recommending Music to Listen to While Working”. In: *Proceedings of the 23rd ACM International Conference on Intelligent User Interfaces*. ACM, 2018, pp. 7–17. DOI: 10.1145/3172944.3172981.