Understanding How People Reason about Aesthetic Evaluations of Artificial Intelligence

Changhoon Oh[1] **Seonghyeon Kim**[2] **Jinhan Choi**[3] **Jinsu Eun**[3] **Soomin Kim**[3]
Juho Kim[4] **Joonhwan Lee**[3] **Bongwon Suh**[3]
[1]Carnegie Mellon University, Pittsburgh, PA, USA [2]NAVER, South Korea
[3]Seoul National University, South Korea [4]KAIST, South Korea
ochangho@andrew.cmu.edu kim.seonghyeon@navercorp.com juhokim@kaist.ac.kr
{jinhanchoi, eunjs71, soominkim, joonhwan, bongwon}@snu.ac.kr

ABSTRACT

Artificial intelligence (AI) algorithms are making remarkable achievements even in creative fields such as aesthetics. However, whether those outside the machine learning (ML) community can sufficiently interpret or agree with their results, especially in such highly subjective domains, is being questioned. In this paper, we try to understand how different user communities reason about AI algorithm results in subjective domains. We designed *AI Mirror*, a research probe that tells users the algorithmically predicted aesthetic scores of photographs. We conducted a user study of the system with 18 participants from three different groups: AI/ML experts, domain experts (photographers), and general public members. They performed tasks consisting of taking photos and reasoning about AI Mirror's prediction algorithm with think-aloud sessions, surveys, and interviews. The results showed the following: (1) Users understood the AI using their own group-specific expertise; (2) Users employed various strategies to close the gap between their judgments and AI predictions over time; (3) The difference between users' thoughts and AI predictions was negatively related with users' perceptions of the AI's interpretability and reasonability. We also discuss design considerations for AI-infused systems in subjective domains.

Author Keywords

Artificial intelligence, aesthetic evaluation, algorithmic experience, neural image assessment, AI in subjective fields

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI);** *User studies;*

INTRODUCTION

The recent advances in machine learning (ML) algorithms have been leading to a greater interest in artificial intelligence (AI) than ever before from not just those in the academic and industrial fields but also the general public. In the areas of computer vision [10, 64], speech recognition [20, 25], and natural language processing [43, 63], the performance of a large number of AI algorithms is already comparable to that of human experts [23, 36, 43]. Additionally, they are expanding into creative areas such as writing [5] painting [21, 61], and composing [9, 29], which have been considered uniquely human activities. AI technology can even be applied to fields that are very subjective and can be interpreted differently by different people, like aesthetic evaluation [65]. Nowadays, these kinds of remarkable developments of AI technology and its various uses are commonly seen through the media.

However, the application of AI algorithms to such creative but highly subjective domains raises questions about whether the various people or groups surrounding them can commonly understand them or sufficiently agree on them. Both subjective domain expertise, as well as AI expertise, could result in entirely different interpretations and understandings of the output of a creative AI algorithm. Current AI algorithm studies, however, are relatively less concerned with differences in perspectives on these things and still focus on improving their performance and producing experimental results. Moreover, AI algorithms sometimes do not fully explain their internal principles, which is sometimes referred to as the black box problem [3, 8, 32]. In situations in which people cannot accept the results of AI algorithms, if the transparency of an algorithm is not ensured, its users may lose confidence in the algorithm and not be immersed in it [24, 60].

With this as a background, we aim to investigate how different users reason about the results of an AI algorithm and discuss human-computer interaction (HCI)/user experience (UX) considerations in the design of AI-infused user interfaces. First, we designed a research probe, *AI Mirror*, a user interface that tells users the algorithmically predicted aesthetic scores of photos based on a deep neural network model (Figure 1). Then, we conducted a user study using both quantitative and qualitative methods. We recruited a total of 18 participants consisting of a well-balanced mix of AI/ML experts, photographers (domain experts), and members of the general public. They performed a series of tasks consisting of taking photos using AI Mirror and reasoning about its algorithm with the think-aloud method and survey. In the survey, we collected users' expected scores for their pictures and their

DIS '20, July 6–10, 2020, Eindhoven, Netherlands.

ACM ISBN 978-1-4503-6974-9/20/07 ...$15.00.
http://dx.doi.org/10.1145/3357236.3395430

interpretability and reasonability ratings for the AI's scores. We also conducted semi-structured interviews about how users experienced the system. The results from the study can be summarized as follows:

- According to their group (i.e., experts, photographers, general public), users showed different characteristics in reasoning about the subjectivity of the AI algorithm. They understood the AI using their own group-specific expertise.
- The group of photographers reported the highest scores in perceived interpretability and reasonability of the AI's aesthetic scores. On the other hand, the AI/ML experts had difficulty interpreting them and considered them relatively unreasonable.
- If there was a difference between the users' thoughts and the AI's predictions, they had difficulty interpreting the AI's predictions and considering them reasonable.
- Users adopted their own personal strategies to infer the AI's principles of evaluation, such as making subtle changes to various picture elements and extending their ideas through various examples.
- While interacting with the AI, users wanted it to be highly interpretable. They wanted to actively communicate with the AI to understand the nature of the subjectivity of its algorithm.

Based on these findings, we discuss design considerations for AI-infused user interfaces that convey subjective results, such as aesthetic evaluations, to users.

The main contributions of this work to the HCI community are as follows:

- A research probe for a black-box-like situation based on neural networks that allows users to experiment with an AI algorithm and develop their own subjective thoughts about it.
- Experimental results showing how the unique characteristics of users affect the process of inferring the outcomes of the AI in a subjective area in terms of group, strategy, and communication.
- Design implications for intelligent user interfaces that deliver a variety of interpretable results, which could be utilized by both the AI/ML and HCI communities.

RELATED WORK

This section addresses the related work of this study with three key topics: (1) Interpretability of AI algorithms; (2) sense-making and gap between users and AI algorithms; and (3) user control in intelligent systems.

Interpretability of AI Algorithms

Despite the remarkable advances of AI algorithms with the development of deep learning (DL), it has been pointed out that it is relatively difficult to understand how the internal principles and mechanisms of the algorithms work [3, 8, 32]. To elucidate the principles of the algorithms, researchers of the AI/ML community have conducted various studies [30, 39, 59, 66]. Some research has been conducted on the topic of explanatory AI (XAI) [1, 13, 19, 22, 45]. As algorithms extend to the domain of human creativity, where people can have various subjective interpretations, the issue of interpretability is becoming even more pronounced.

The HCI community has also regarded algorithms as an important research topic [14, 16, 38, 48]. In particular, many studies have focused on the fairness and transparency of algorithms [18, 26, 35, 54, 60, 68] along with their interpretability. Some studies suggest that algorithms could often be less objective than required, increasing bias [6, 50]. Users' concerns about the bias and opacity of algorithms can potentially affect their trust in the user interfaces that operate on them as well as the algorithms themselves [15, 69].

Beyond the individual user level, the problem of algorithms can be extended to issues of various groups surrounding AI technology [69]. AI technology involves a wide range of stakeholders [12], not just technical experts but experts in a variety of specialized domains and the general public who could potentially use the technology. In this sense, when planning or creating AI-infused services or products, it is necessary to consider various stakeholders rather than simply taking a user-centric perspective [53]. This becomes even more important in relation to the expansion of AI technology into not only simple and repetitive tasks but also subjective domains that can be interpreted in diverse ways by the various groups involved.

Sense-making and Gap between Users and AI Algorithms

Sense-making is a set of processes initiated when people recognize the inadequacy (gap) of their current understanding of events [31, 55]. In this situation, individuals build, verify, and modify their mental models to account for the unrecognized features. Since the concept has been considered a framework to understand the interaction between people and information technology [33, 42, 47, 56], numerous studies have used it as a research method [11, 44] or introduced interactive systems for supporting it [57, 62].

The concept and framework could also be applied to the understanding of how people reason about the results of AI algorithms. As AI algorithms are producing and communicating results that go beyond what people can understand, there could be differences between the results of AI and human perceptions, especially in subjective domains. Looking at the processes people use to reduce the differences between their thinking and the results of AI algorithms can provide important information about how people interact with AI algorithms and AI-infused interfaces.

User Control in Intelligent Systems

In the HCI community, there has been discussion of how users and automated systems communicate [58]. Some have conducted research based on the idea of developing an adaptive and intelligent agent that automatically responds to user behavior [28, 40]. In contrast, other groups of scholars have argued that a system encouraging users' ability to manipulate interfaces directly should be considered [2]. In addition, a mixed-initiative viewpoint has been raised that combines the two to take advantage of each [4, 17, 27].

Recent advances in AI algorithms have rekindled interest in these discussions, since AI algorithms can now respond to user behavior more intelligently than ever before and users are communicating in new ways rather than simply manipulating the interfaces. In particular, as AI is used in subjective fields that can be interpreted in various ways, multi-faceted analysis, in-depth understanding, and independent implications on the interaction between users and AI in the field are required. In this study, we closely observe and analyze the interaction between the user and AI algorithms to see if control and communication could provide value to users and extend this discussion.

Research Questions

Based on this background, we would like to address the following research questions in our paper.

- Can users regard the results of AI in subjective domains as interpretable? Can they feel that the judgments made by AI algorithms are reasonable?
- What difference does expertise (domain expertise and AI/ML expertise) make when users interact with AI?
- Can users narrow the difference between their thoughts and AI predictions through constant interaction with AI?
- How do users want to control and communicate with AI algorithms that provide information on subjective domains?

AI MIRROR

To address the research questions, we designed a research tool, AI Mirror.

Design Goal

In the design of the research tool, we aimed to create an interface that provides users with a black-box-like situation where they can interact with AI algorithms but do not know exactly how they work. In order to answer our research question, the domain being explored had to fit a couple of parameters: (1) introducing state-of-the-art neural network algorithms that provide results but do not provide details about the calculation process and (2) selecting a topic that allows users to produce their own artifacts and interpret the results of AI on them. Among the creative and open-ended domains, we selected aesthetics. We reviewed *Augury* [34], which evaluates a website's design by calculating the complexity and colorfulness of the website with a database of aesthetic preferences, and used the concept in the design process. Finally, we created an interface that can predict the aesthetic quality of photographs provided by users based on a state-of-the-art neural network algorithm and named it "AI Mirror" (Figure 1).

Image Assessment Algorithm

In the design of AI Mirror, we utilized Google's Neural Image Assessment (NIMA) [65], an AI algorithm to predict the aesthetic quality of images. This convolutional neural network (CNN) is trained to predict which images a typical user would rate as looking both technically good and aesthetically attractive. Specifically the algorithm trained both the AVA dataset [46] and the TID2013 dataset [52]. The former contains about 255,000 images, rated based on aesthetic qualities by amateur photographers and the latter contains 3,000 of test images obtained from 25 reference images, 24 types of distortions for each reference image, and 5 levels for each type of distortion. Instead of classifying images according to low/high scores or regressing to the mean score, the NIMA model produces a distribution of ratings for any given image on a scale of 1 to 10 [65]. In the process of creating AI Mirror, we refined it through repeated ideation and revision. we identified that as the mean scores of given images approximated the normal distribution, the scores concentrated on the average, and extreme values were rarely found. Since it was possible that users could not perceive the difference between the good and bad pictures, we performed a linear transformation of the normal distribution so that users fully utilized the algorithm in the experiment.

Design of User Interface

AI Mirror was developed as a web application that works on a mobile web browser and uses a camera and photo album. The user interface of AI Mirror is composed of four main views:

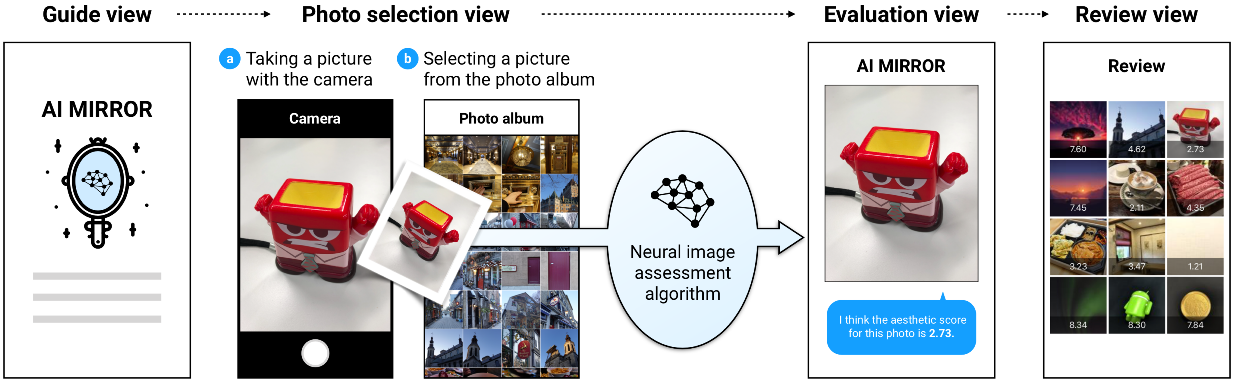
Guide view
Photo selection view
Evaluation view
Review view
AI MIRROR
a Taking a picture with the camera
b Selecting a picture from the photo album
Camera
Photo album
Neural image assessment algorithm
AI MIRROR
I think the aesthetic score for this photo is 2.73.
Review
7.60
4.62
2.73
7.45
2.11
4.35
3.23
3.47
1.21
8.34
8.30
7.84

Figure 1. We designed a research probe, "AI Mirror," a user interface that tells users the algorithmically predicted aesthetic scores of photos based on a deep neural network model. To investigate how different users reason about the results of an AI algorithm, we conducted a user study of the system with 18 participants from three different groups: AI/ML experts, domain experts (photographers), and general public members.

guide view, photo selection view, evaluation view, and review view (Figure 1).

- Guide view: This screen is the first screen of the user interface and provides a basic explanation of the system with a simple concept image. A user can start using the system by entering his or her username.
- Photo selection view: On the next screen, the user can select a picture to be evaluated by the AI. In the process, the user can select one of two functions: (1) taking a picture with the camera app and (2) selecting a picture from the photo album. If the user chooses the former, the camera of the smartphone is activated so that the user can take a picture. If the user selects the latter, the photo album is activated so that the user can browse pictures and select a picture.
- Evaluation view: Right after taking or choosing a picture, AI Mirror shows the user the aesthetic score of the picture on a 10-point scale, stating in text, "I think the aesthetic score for this photo is 8.99."
- Review view: In this view, the user can view the pictures that have been evaluated by the AI so far. Photos are presented in a tile layout. If the user selects a photo, the photo is enlarged in the pop-up window. The aesthetic score is displayed at the bottom of the photo.

STUDY DESIGN

To understand how users interact with the system, we designed a user study with a mixed-methods approach using both quantitative and qualitative methods.

Participant Recruitment

In recruiting participants, we sought to balance the following three groups: AI/ML experts, photographers, and the general public. We set specific recruitment criteria for each group. First, the AI/ML experts group included only those who had majored in computer science-related areas, such as ML and DL, or had experience as specialists in related fields. The group of photographers included professional photographers, people with training in photography, or non-professional photographers who had more than 10 years of photography experience. In particular, each group included only those without expertise in the area(s) of the other expert group(s). For the general public group, people who did not have these types of expertise were sought. We first posted a recruiting document on our institution's online community and then used the snowballing sampling method. We recruited a total of 18 participants, with the same number of participants for each group (Table 1). Some of the recruited AI experts ran AI-related startups, some had authored papers for top conferences in computer vision, while others were involved in related industries with relevant knowledge and expertise. The recruited photographers were people with more than 10 years of photography experience.

Experimental Settings

In the user study, we used a dedicated device, the iPhone X, as the main apparatus to control the experiment by providing the same conditions to all participants.

ID	Sex	Age	Characteristics
A1	M	37	CTO of video AI startup (author of CVPR paper)
A2	M	35	CEO of video AI startup (author of CVPR paper)
A3	M	32	CTO of sound AI startup (author of DCASE paper)
A4	M	29	AI Researcher (teaching ML/DL experience)
A5	F	26	AI Researcher at IT center (majoring in ML)
A6	F	24	AI Researcher at IT company (AI field strategy)
P1	F	34	Photographer (10 years of field experience)
P2	F	30	Amateur photographer (took camera course)
P3	M	34	Photographer (10 years of field experience)
P4	F	31	Photographer (10 years of field experience)
P5	M	30	Amateur photographer (11 years of experience)
P6	F	29	Photographer (majoring in fine arts)
N1	F	36	Administrative worker
N2	M	34	Researcher (urban planning)
N3	F	25	Nursing teacher
N4	F	32	Graduate school student (communication studies)
N5	M	28	English teacher
N6	M	28	Graduate school student (business)

Table 1. Participant information. (IDs: "A" = AI/ML expert, "P" = photography expert, "N" = no expertise (general public).)

Since the experiment was done in the laboratory, it was necessary to prepare a variety of objects and additional material that participants could use to take pictures. Various objects of different colors and shapes (e.g., a yellow duck coin bank, a green android robot figure, a blue tissue box) were prepared so that users could combine various colors and attempt various compositions with them. In addition, we prepared backdrops so that users could keep the background clean and clear. We also prepared a small stand light. This setup allowed participants to freely take various photos. Meanwhile, to satisfy the users' various photo selection needs, we entered many pictures in the photo album of the experimental device beforehand. This photo album consisted of pictures that had been evaluated by AI Mirror in advance. We selected an equal number of images (10 images) in eight sections scored from 1 to 8 points, and finally, 80 images were included.

Procedure

In the study, participants completed a series of tasks, including interacting with AI Mirror and reasoning about its algorithm with the think-aloud method and survey, and then took part in interviews. In a separate guidance document of AI Mirror before the experiment started, we provided the participants with a detailed explanation of the purpose and procedure of the experiment. Users were allowed to manipulate the system for a while to get used to it. On average, the experiments lasted about 60 minutes. All participant received a gift voucher worth $10 for their participation.

Task

The main task that participants were asked to perform in the experiment was to interact with AI Mirror and deduce the photo evaluation criteria of AI Mirror. Using AI Mirror, the participants took photos or selected photos from the photo album, and AI Mirror made aesthetic evaluations of the photos. There were no particular restrictions on the number of interaction trials or experiment times.

Survey

Each time a picture was taken or selected, participants were asked to respond to the survey. In each trial, we asked participants to answer three questions. The first asked participants about the expected score for the aesthetic evaluation of the photo they had just taken or selected. Just prior to AI Mirror's aesthetic evaluation of the picture, they were asked to give their own score on a 10-point scale. The second question asked participants whether they found AI Mirror's aesthetic evaluation score interpretable. Participants rated this on a 5-point Likert scale (1=strongly disagree, 5=strongly agree). The third question asked participants whether AI Mirror's aesthetic evaluation score was reasonable. Likewise, participants rated this on a 5-point Likert scale (1=strongly disagree, 5=strongly agree). In addition to these three items–*expected score*, *perceived interpretability*, and *perceived reasonability*– we measured the *difference* between the participant's aesthetic score *(expected score)* of the picture and that of AI Mirror. We also collected the *trial number* (e.g. trial 1, trial 2) along with these values to capture temporal changes in values for each participant.

Think-aloud Session and Interview

We conducted a qualitative study using the think-aloud method [67] and semi-structured interviews to gain a deeper and more detailed understanding of users' thoughts. While performing the tasks, the participants could freely express their thoughts about the tasks in real time. We audio recorded all the think-aloud sessions.

After all tasks were completed, we conducted semi-structured interviews. In the interviews, the participants were asked about their overall impressions of AI Mirror, especially focusing on its interpretability and reasonability. All the interviews were audio recorded.

Analysis Methods

From the study, we were able to gather two kinds of data: quantitative data from the surveys and the system logs and qualitative data from the think-aloud sessions and interviews.

Quantitative Analysis

In the quantitative analysis, we conducted statistical analysis of the six types of data collected from each participant (*group, trial number (trials), expected score, difference, perceived interpretability,* and *perceived reasonability*) and tried to identify significant relationships among and differences between these variables. We used panel analysis as the main analysis method, since it is specialized for analyzing multidimensional (cross-sectional)" data collected "over time (time-series)," and from the same individuals, which matched the data we gained from our experiment exactly. Besides, it can run the regression model of each variable of multidimensional data, so it can provide more concise and comprehensive results than an ANOVA, which produces results in an aggregated way without considering the time effect. Moreover, as we recruited users by group, we assumed that the unobserved variables were uncorrelated with all the observed variables and accordingly used a random effects model.

According to our RQs, we selected *difference*, *perceived interpretability*, and *perceived reasonability* as dependent variables and made regression models for each DV. The regression models (1) *difference* $\sim$ *trials* (from *trial number*) + *group*, (2) *perceived interpretability* $\sim$ *trials* + *difference* + *group*, and (3) *perceived reasonability* $\sim$ *trials* + *difference* + *group* are presented in Tables 2, 3, and 4, respectively, with brief summaries of r-square, f-score, and significance levels. Although *difference* was used as the independent and dependent variable depending on the model, it was not used as the IV and DV in any one model at the same time.

Qualitative Analysis

The qualitative data from the think-aloud sessions and post-hoc interviews were transcribed, and analyzed using thematic analysis [7]. In the process, we used Reframer [70], a qualitative research software tool provided by Optimal Workshop. To organize and conceptualize the main themes, three researchers used line-by-line open coding. Through a collaborative, iterative process, we revised these categories to agreement and then used axial coding to extract the relationships between the themes.

RESULT 1: QUANTITATIVE ANALYSIS

In the case of the statistical analysis results, the emphasis was on understanding the basic relationship or tendency between variables. As described above, in each user's trial, *trial number, difference*, user's *perceived interpretability* and *perceived reasonability* evaluations of the AI score, and the user's *group* were collected. Since we did not set any particular restrictions on the number of interaction trials, the number of trials among the participants was slightly varied, with a minimum of 10 and a maximum of 27 (M=14.22, SD=4.48). From 18 participants, we collected a total of 256 data points for the same set of variables, which we considered large enough to conduct panel analysis. Through the analysis on each DV (*difference*, *perceived interpretability*, and *perceived reasonability*), we were able to identify significant relations between some of the variables. We report the results for each regression model in order. Statistically significant results are reported as follows: *p<0.001(***), p<0.01(**), p<0.05(*).*

Difference

First, in the analysis on *difference*, based on the results shown in Table 2, we observed that *trials* had a significant influence on *difference* (t-value=-2.66, p<0.01**). That is, as the number of trials increased, *difference* significantly decreased, which means that as users continued to interact with the AI, they reduced the difference between their expected scores and the AI's scores. In addition, although we did not identify any significant effects of *group*, we found that there were slight differences in *difference* between user groups. Surprisingly, AI/ML experts showed the biggest difference from the AI (Mean=2.14), followed by the general group (Mean=2.07), and finally the photographers (Mean=1.84).

Variable	$\hat{\beta}$	*Std. Error*	*t-value*	*p-value*	
(Intercept)	2.611	0.238	10.970	<0.001	***
Trials	-0.049	0.018	-2.656	0.008	**
General public	-0.203	0.244	-0.830	0.407	
Photographer	-0.373	0.236	-1.582	0.115	

R^2=0.035, Adj. R^2=0.023, F(3, 252)=3.02, p-value=0.03*.

Table 2. Results of panel data analysis of *difference* (*difference* ∼ *trials* (from *trial number*) + *group*). Baseline (Intercept) represents the condition of *group* of AI/ML experts.

Perceived Interpretability

In the analysis on *perceived interpretability*, based on the results shown in Table 3, we observed that *difference* had a significant influence on *perceived interpretability* (t-value=-7.63, p<0.001***). That is, as *difference* increased, *perceived interpretability* significantly decreased, which means that users had difficulty interpreting AI scores when there was a big difference between their evaluations and those of the AI. In addition, we identified that *group* had a significant effect on *perceived interpretability*, especially for photographers (t-value=4.86, p<0.001***). The photographer group (Mean=3.90 out of 5) showed a higher level of interpretation of the aesthetic scores evaluated by the AI compared to the AI/ML experts (Mean=2.44). Although it was not a significant difference, the general public (Mean=2.96) also showed a higher level of interpretation than the AI/ML experts. Meanwhile, *trials* also showed a slightly positive effect on *perceived interpretability*, but it was not significant either.

Variable	$\hat{\beta}$	*Std. Error*	*t-value*	*p-value*	
(Intercept)	3.034	0.251	12.090	<0.001	***
Trials	0.017	0.014	1.280	0.202	
Difference	-0.334	0.044	-7.615	<0.001	***
General public	0.488	0.274	1.782	0.076	
Photographer	1.323	0.272	4.860	<0.001	***

R^2=0.280, Adj. R^2=0.268, F(4, 251)=24.36, p-value<0.001***.

Table 3. Panel data analysis of *perceived interpretability* (*perceived interpretability* ∼ *trials* + *difference* + *group*). Baseline (Intercept) represents the condition of *group* of AI/ML experts.

Perceived Reasonability

In the analysis of *perceived reasonability*, based on the results shown in Table 4, we observed that *difference* had a significant influence on *perceived reasonability* (t-value=-12.02, p<0.001***). That is, as *difference* increased, *perceived reasonability* significantly decreased, which means that users did not think the AI score was reasonable when there was a difference between their thoughts and those of the AI. We also identified that *group* had a significant effect on *perceived reasonability* in the case of photographers (t-value=3.33, p<0.001***). The photographers gave higher reasonability scores (Mean=3.74 out of 5) than AI/ML experts did (Mean=2.43). Although it was not a significant difference, the general public also gave higher reasonability scores (Mean=2.92) than AI/ML experts did. On the other hand, *trials* slightly lowered the *perceived reasonability*, but it did not show any significant effect.

Variable	$\hat{\beta}$	*Std. Error*	*t-value*	*p-value*	
(Intercept)	3.797	0.268	14.156	<0.001	***
Trials	-0.024	0.013	-1.884	0.061	
Difference	-0.485	0.040	-12.021	<0.001	***
General public	0.277	0.318	0.872	0.384	
Photographer	1.057	0.317	3.334	<0.001	***

R^2=0.410, Adj. R^2=0.401, F(4, 251)=43.63, p-value<0.001***.

Table 4. Results of panel data analysis of *perceived reasonability* (*perceived reasonability* ∼ *trials* + *difference* + *group*). Baseline (Intercept) represents the condition of *group* of AI/ML experts.

To summarize the results of the quantitative analysis, first, we partially identified that users in different groups showed differences in the process of interacting with the AI. The group of photographers showed the highest perceived interpretability and reasonability among the three groups, with AI experts having the lowest. Second, users were able to narrow the gap between their evaluation scores and those of the AI as they continually interacted with AI. Third, higher difference between users' thoughts and the AI's predictions lowered both the perceived interpretability and reasonability of the AI.

RESULT 2: QUALITATIVE ANALYSIS

In the qualitative analysis results, we focused on finding detailed features not revealed in the statistical analysis. Here, we report the characteristics of each group, the strategies users showed in the reasoning process, and the factors users considered important in their interpretability and reasonability evaluations. (The quotes are translated into English.)

People Understand AI Based on What They Know

Through the qualitative analysis, we identified that while interacting with AI Mirror, the participants showed distinctive characteristics according to their group. In particular, we observed that the vocabulary they used reflected their expertise. Each participant also attempted a distinct approach in the process of reasoning.

First, while interpreting the AI's results, AI/ML experts commonly used words that reflected specialized knowledge of ML and DL, such as "algorithm," "dataset," "training," "model," "black box," "pixel," "classification," and "feature," which were never mentioned by the other groups. For example, A1 said, *"It's like evaluating a model. It's like putting unseen data into the test set and seeing if it works or not."* A6 said, *"There may be some problems with the learning process and the database. It depends on if it was based on social media data, like Instagram. You know, colorful photos usually get a lot of likes."* A5 said, *"And I think we should open the black box if possible and make it a white box."* People in this group also used their AI/ML expertise in inferring the AI's criteria. For instance, A4 said, *"I think the boundaries of this object are not clear. It seems the algorithm is not detecting this object well. Normally vision technology needs to know the boundaries of objects."* A4 then edited the photo of a white egg with a white background by drawing the outline of the egg. However, unexpectedly, AI/ML experts did not receive high scores overall and eventually said they were not confident in their understanding of the AI's standards. In browsing the pictures that he took on the review view of AI Mirror, A2 said, *"I do not know why this score is high .... and this is too low a score."* A3 concluded the experiment by saying, *"The experiment itself is interesting .... but my pictures scored much lower than I expected."*

Secondly, the photographers interpreted and inferred the results of the AI using their expertise in photography. They often mentioned important elements of photography, such as "light," "color," "moment," "composition," and "distance," and camera controls, such as "focus," "aperture," and "lens." For example, P2 said, *"This picture has a low depth of field, so I think it will get a higher score than the previous one."* P4 said, *"The composition of this picture follows the rule of thirds well."* P1 said, *"The light is concentrated toward the black background, so this doll is too bright. So I'm going to adjust the light by touching it on the camera app screen. I often do this. This makes the background darker and darker."* When choosing images in photo albums, people in this group also picked the pictures that seemed likely to get high scores from the AI, taking advantage of their expertise. Taking the viewpoints of the photographer of the picture that he picked from the album, P3 said, *"This is definitely a good picture. The photographer must be proud of such a beautiful picture. He must have waited for this moment."* Emphasizing the importance of photoshopography, P6 also said, *"I think this photographer did photoshopography on this image to express the colors of various spectrums."* P5 also assessed the quality of the photo selected from the album by reasoning about the weather at the time of the picture. Overall, the group of photographers took or picked high-scoring pictures, often showing expected scores similar to those of the AI. When he realized that his score was almost identical to the AI's score, A1 was surprised, and he said, *"It was creepy. I think this AI is in my head."*

Third, the general public group took pictures in the way that they typically take pictures without specific professional knowledge. They mainly took pictures of their favorite objects from among those prepared for the experiment or chose pictures of famous landmarks or beautiful landscapes from the photo album, believing that the AI would appreciate these pictures. For example, N1 said, *"This [a yellow duck coin bank] is really cute. I'll take this."* N3 said, *"I'm just looking for a picture that looks pretty. This picture is pretty. Everything in the picture looks pretty. It looks like a house from a fairy tale."* Looking at the photos in the photo album, N6 said, *"And I think I've seen this quite a few times. It's the Louvre Museum,"* and picked the photo. However, they did not fully comprehend the scores of the AI. N2 said, *"I think there must be a certain standard ... But I cannot quite grasp it. I do not know if it's really aesthetic judgment."* N5 said, *"I think the AI has another evaluation criterion. The AI does not think this picture is pretty."* N4 even complained, saying, *"I think it'll give a very high score to this picture. Actually, I do not think this picture is pretty. However, the AI has always been so contrary to me, so this picture will have a high score."*

People Reduce Differences Using Various Strategies

Next, we identified that as they continued to interact with AI, participants adopted their own personal strategies to infer the AI's principles of evaluation. They used approaches that involved making subtle changes to various picture elements, and they extended their ideas through various examples.

First, when participants took pictures, they tried to experiment with the AI by making slight changes to the pictures. They changed the background color of an object or the composition of the same object. They sometimes added objects one at a time and looked at the AI's reactions as different colors were added. P1 said, *"The next thing I wanted to do was keep the background white and add another colored object. I wanted to see how the score changed when I did that."* N6 said, *"This time, I'll take the same background and object from a distance. It makes the object look small in the picture. I have to change only one element .... Oh .... 4.75 points. Size does not matter. Now I understand more."* N3 said, *"And this time, I'll take this same object on a yellow background. I think if the background is yellow, somehow it looks like the background will be more in focus than the object, so the score will be lower. (Score: 2.19) Now I know more. I think the AI and I have a similar idea."* Through this process, most of the users found that the AI gave high scores (8 points on a 10-point scale) when one bright object in the photo stood out against a black background. Photographers tried these kinds of pictures relatively earlier in their trials than the other participants did.

Second, some participants even used the editing features of the iPhone photo app to actively modify the photos they took or the photos they picked from the album and asked the AI to evaluate the modified photos. A4 described, *"I'll edit this photo of the macaroons. Let me get rid of the color. The reason for doing this is to know if the color is important or not. The color has gone and I think it will be lower than 7.22."* P5 said, *"I'll crop the photo. Let's move the object to the center. I just changed the position of the object. I think this picture will be rated at about 8 points. (Score release) Uh-oh (...) The score is lowered (...) The composition is not a problem."* In this way, participants developed a better understanding of the characteristics of AI by creating slightly different versions of the photographs. They all stated that this process enabled them to better understand and experience AI principles.

Third, participants applied their speculations about how AI works to different cases. They continued their testing of the aesthetic evaluation criteria of AI by using similar examples. They wanted to know whether the criteria they had grasped could be applied to other photos with similar characteristics (e.g., composition, contrast, color) but different objects from the photos they had taken or identified so far. N2 explained, *"I'll pick this crab signboard picture. I think this is going to have a score similar to the picture I took before. What was the score of the photo with the white background and the red toy?"* A5 described, *"I'll pick a photo with a variety of objects and a central object in it. That's the standard I've figured out so far."* After getting a high score for a photo with a black background, P1 said, *"Then, this time, I'll pick a picture with a black background similar to the last one."* Through this process, participants were able to confirm whether their criteria were correct and narrow the gap between their thoughts and those of the AI.

Lastly, we identified that participants tried to find new standards that they had not seen so far by choosing completely different pictures from the photo album. After finding a certain way to get a high score, some participants additionally attempted to look at new types of photos. P4 described, *"I'll take a look at the kinds of pictures I have not seen before. I'll try this .... I have to review the various pictures to see what it likes and what it does not like."* P3 said, *"I'll try it again. Um .... I'll take this. This is just a pattern that I have not picked up so far."* N3 also remarked, *"I just want to try something I have not tried yet. I think it likes pictures of things that are distinct and colored. But from now on, I do not think I should choose things like that."* Through this process, participants were able to find new and unexpected criteria, such as *"a preference for photographs with repetitive patterns."*

Overall, based on these various strategies, while interacting with AI Mirror, participants were able to understand its scoring system and narrow the gap between its scores and their scores.

People Want to Actively Communicate with AI

Finally, regarding users' perceptions on the interpretability and reasonability of the AI algorithm's aesthetic evaluations, the participants wanted to actively communicate with AI Mirror in the experiment.

During the think-aloud sessions and interviews, regardless of their group, participants recounted interacting with the AI as a positive experience. Most participants described the process as interesting, fun, and enjoyable. In particular, while reasoning about the criteria of AI Mirror's aesthetic evaluations, participants felt curious about the principles of AI and wanted to know about it. P4 described, *"It was fun and interesting. It got me thinking. It stimulated my curiosity."* N1 expressed, *"It was fun to find out the criteria it used to rate them. It was just an experiment, but I was really curious."* Participants were also delighted when the difference between the AI score and their expected score was not that large. They were even more delighted when the AI gave a higher score than they expected. They expressed that it was as if AI Mirror had read their thoughts and that they felt like they were being recognized and praised by the AI. N3 said, *"Later, I felt good about the AI, because it was well aware of the points I had intended and appreciated my effort."* N5 said, *"I feel good because I got a high score. I feel like I'm being praised by the AI."* Some participants even asked us to send the URL link to the AI Mirror webpage at the end of the experiment. They wanted to get ratings on their personal smartphone photos and to interact more with the AI.

Nonetheless, most participants stated that they also felt negative emotions during the interaction. When their expected scores differed significantly from those of the AI, especially when they were rated very poorly by the AI, participants felt embarrassed, unhappy, and frustrated. For example, N5 described, *"Oh .... I feel terrible. This score is lower than the previous one. I took more care with it. I feel worse as my score drops. It's pretty unpleasant."* Participants told us that they could not understand why the AI's scores were lower than they thought and that they had difficulty interpreting the results. N6 said, *"I'm so frustrated because I do not know why my score is so low."* A2 complained, saying, *"This is really low, but I do not know why .... This is too low .... I know this is an ugly picture. But is there a big difference from the photo I took earlier? (His previous picture scored very high)."* Some even expressed that they could not understand the AI and regarded this interaction as meaningless. P6 said, *"Maybe it just thinks so. It is just being like that. I do not want to deduce anything. My overall level of interest is .... pretty low. I have no understanding of it."* These unpleasant experiences also reduced participants' trust in the system as well as their confidence that they could take pictures well. P2 said, *"I think this picture will get 6 points (She gradually scored lower and lower on her photographs). I have lost my confidence. I think my expectations for my picture have been lowered too."*

In such a situation, the absence of communication between users and AI can be considered the main cause of the negative emotions of users. During the interviews, participants uniformly expressed a desire to communicate with the AI. They wanted the AI to explain not only the calculated scores but also the detailed reasons. N6 said, *"I wanted to know the elements of the scores. I think it would be better if it could tell me more specifically."* P6 expressed, *"It would be much better if it could tell me why it came up with this score. Then I could take better pictures."* Furthermore, participants wanted to let the AI know their thoughts. P4 said, *"I want to let the AI know this is not as good a picture as it thinks."* A6 described, *"I had a lot of disagreements with the AI. I think it would be nice if it could learn my thoughts on the points on which we disagreed. It is my app, and it has to appreciate what I think is beautiful."* Some participants said that in this one-sided relationship, even though they could interpret the evaluations of the AI, they could not see them as reasonable. P1 said, *"The weather in the photos is not that sunny, but I like the cloudy weather. I'm sure that AI Mirror will rate this picture too low. It only likes those pictures that are high contrast. I can clearly see why the score is low, but I cannot say that it is reasonable."*

The various emotions that the participants experienced during the user study and their strong desire for communication for improved interpretability and reasonability suggest that in the design of user interfaces with AI (namely, algorithms), additional and careful discussion is needed.

DISCUSSION

In this section, we discuss lessons learned from the user study and its implications for AI-infused user interfaces conveying subjective information that can be interpreted in diverse ways. We also report our plans for future work as well as the limitations of the study.

Different Perspectives on AI Algorithms

Through the user study, we identified that users interpreted AI in different ways according to their group (result 1 from the qualitative analysis). AI/ML experts tried to find out the characteristics of its training data and learning process based on their knowledge of AI. Photographers looked at it considering the elements of photography and cameras. The non-experts tried to understand it based on their impressions of the photos without relevant prior knowledge. Most notably, contrary to our expectations, AI/ML experts showed the greatest difference from AI and the lowest interpretability and reasonability scores (Tables 3 and 4). On the other hand, the photographers showed the smallest difference from AI and the highest interpretability and reasonability scores (Tables 3 and 4).

The results raise the question of why domain experts (photographers) rather than AI/ML experts interpreted the AI algorithm the best and narrowed the gap with it the most. One explanation could be that the training dataset used in the model construction reflected, to some degree, the domain expertise. The NIMA algorithm [65] that was infused in AI Mirror was built based on both the AVA datasets [46] and the TID2013 dataset [52]. The former includes amateur photographers' 10-point-scale aesthetic evaluation ratings on 250,000 photos from an online photography community. The latter contains images collected through a forced choice experiment, where observers select the better image between two distorted choices of reference images. In the process of distorting them, the elements considered important in the photographs, such as compression artifacts, noise, blur, and color artifacts, were reflected, which were often mentioned by domain experts in the experiment. Through such a process, the knowledge of the domain expert could be incorporated into the training data, and the generated model might have had a view more similar to that of the domain expert group than those of other groups.

Meanwhile, from an AI/ML expert's point of view, there are some points worth mentioning. They may have had biases from their existing knowledge on ML algorithms to which they were accustomed. In the experiment, we did not explain anything about the model infused in the prototype, so even the members of the AI/ML experts group did not know which algorithm was working behind it. They seemed to have trouble because they tried to engage in sense-making by incorporating the techniques with which they were familiar, which could have been different from the actual principle. Rather, the general public, who had no prior knowledge, could look at the picture without a biased perspective, and this might have helped them to show results that were more similar to those of the algorithm than those of ML experts.

These results suggest that it is essential to consider what expertise is reflected in the training dataset of the model behind the AI system. Understanding who annotated the data and what criteria were reflected can be as important as building a model and improving its performance. Accordingly, AI professionals need to actively communicate with domain experts of the algorithms they are trying to build, and of course, domain experts need to be actively involved in this process. Furthermore, in developing such interfaces, it is necessary to understand and reflect not only the specialized domain but also the user's point of view and provide appropriate information regarding it. Especially in subjective areas like aesthetics, providing information to the user, such as what data it uses to make decisions and whose views are reflected, could be a helpful way for users to understand how the system works and appreciate the results.

- **Implication:** Consider what domain view the AI system's algorithm in the subjective domain contains-who annotated what data and what criteria were reflected-and provide the user with the information based on it.

Users' Strategies to Reduce Gap with AI

In the absence of any information about AI Mirror's aesthetic score calculation process, users were curious about the algorithm and constantly strived to learn the principles actively through various strategies (result 2 from the qualitative analysis). Sometimes, they formulated hypotheses and tested them by taking slightly different photos. Other times, they just explored without a clear hypothesis or direction. Through these strategies, they were finally able to narrow the gap between their thoughts and those of AI (Table 2). We can think of design implications on both the user side and the AI side.

First, on the user side, we can consider introducing these factors into the design of tools that help people to understand AI/ML models, which has recently received a great deal of attention [51]. AI can use the strategies people utilized to help them understand its principles. An AI interface needs to prepare and show as many examples as possible so that people can understand the principle as easily as possible. It is also possible to improve the user's understanding in a microscopic manner by preparing several examples and images with small but clear differences. A potential macroscopic approach is presenting users with completely different examples or images to enable new and diverse ideas and expand their thinking. However, presenting various examples to users may not be enough. It might be helpful to give examples with explanations that the model can be complex and that it will not be judged by simple rules to make the mental model of the user's system clearer. Through these, the public would be able to reduce the differences between their thoughts and those of AI and understand the principles of algorithms easily.

On the other hand, the various strategies and willingness to discover the principles shown by users suggest implications for the AI domain in relation to the production and securing of high-quality data. According to information gap theory, when people are made aware of a gap between what they know and what they would like to know, they become curious and engage in information-seeking behavior to complete their understanding and resolve the uncertainty [41]. This innate desire to satisfy their curiosity can be helpful in gathering information about the way users interpret AI. Through this, we might collect feedback on various use cases and utilize it to improve algorithms. Indeed, curiosity interventions have been shown to improve crowd worker retention without degrading performance [37].

Designing a platform for AI to stimulate users' curiosity and receive various opinions would be useful in securing the large-scale, high-quality data necessary for algorithm refinement learning. AI should be designed to learn from users and narrow the gap with them rather than users adapting to AI. Such a view might have implications for online learning or reinforcement learning generally, as systems can adaptively learn from user feedback and thereby improve themselves.

- **Implication:** Provide interactions with a variety of examples and appropriate explanations so that users can fully understand and appreciate the principles of AI. User interactions, in turn, can be an important source of quality data, especially in subjective areas.

Communication between Users and AI

Finally, we focus on communication between users and AI in subjective fields. Although users tried various ways to understand the AI, they eventually expressed great dissatisfaction with the lack of direct communication with the AI (result 3 from the qualitative analysis). Users wanted the AI to give them more detailed descriptions directly, but they also wanted to explain their ideas to the AI. Some users even felt negative feelings and sometimes lost confidence in the absence of such communication.

In particular, since aesthetic evaluation is intrinsically highly subjective, the problem of communication due to this difference in interpretation may be even more significant. Many AI algorithms have taken a kind of one-way approach, such as recognizing and classifying objects, calculating and predicting values, or improving their performance in those processes. For them, it may be more essential to convey the information to users accurately than anything else. Then, more intelligent interfaces using advanced algorithms have been introduced, but they have adopted a system-driven "adaptive approach." They have tried to understand the user's context and reduce the user's repetitive behavior. In creative and subjective fields such as aesthetics, however, while it is important for AI to provide a convenient path to the user, it is also necessary to understand the user's complex ideas and actively accept their opinions. Developing a creative idea is not a task to be done alone; it also requires a person from whom to seek thoughts or feedback and a process of expressing one's own ideas to persuade and understand the person.

In designing AI-infused systems in creative and subjective fields, we can consider a mixed-initiative approach [27], where users as well as the system can drive the interaction. Introducing communication channels for users and algorithms in the design of AI-based interfaces would be one way of doing this. On the one hand, the AI needs to present users with detailed explanations of the reasons for its decisions [1, 19, 22, 27, 49], showing that its decisions are not arbitrary but based on logic. It may be a good idea to provide specific figures, such as the basis for such a decision and the degree of reliability (i.e., confidence level) of the result. Informing the user at an appropriate time of the circumstances under which the system works better or the conditions that can cause errors can help the user to understand the system and maintain trust.

On the other hand, users should also be able to present their opinions to the AI. In the middle of the interaction with the AI, procedures need to be included that allow the user to express his or her thoughts and enter them in the system. The AI should be able to accept a user's opinion, take it as a dataset, and reflect it in the learning process of the algorithm. Rather than a static AI that only presents predetermined results, a dynamic and adaptable AI that responds to users' thoughts and controls would be more desirable for subjective fields.

- **Implication:** Especially in AI-infused systems in subjective and creative realms, it is very important for users to develop their ideas and reflect them in the system. The system should be designed so that it is not static but dynamic and users and AI can adapt to and understand each other through two-way communication.

Limitations and Future Work

There are several limitations of this study. First, in the questionnaire analysis, the explanatory power of the model was relatively low, although several significant relationships and differences were found. The reason seems to be that the number of participants and trials was too small due to limitations of the experimental environment. Second, we assumed a one-sided relationship between users and AI and did not measure the effect of users' various communications with AI. Third, although we built a black-box-like situation to create a more immersive environment for participants, it made the data we collected and the results of analysis dependent on the participants' inferences and post evaluations.

In future work, we aim to determine a clearer relationship between the various variables by carrying out an expanded study with more participants. We also plan to conduct a user study where participants can experience the AI system in a real context rather than in a controlled environment. In addition, we plan to improve our research tool to cover various areas of AI rather than limiting it to aesthetic evaluation. Finally, we will conduct research that demonstrates the practical effects of the design recommendations that we have proposed.

CONCLUSION

In this study, we investigated how users reason about the results of an AI algorithm, mainly focusing on their interpretability and reasonability issues. We designed AI Mirror, an interface that tells users the algorithmically predicted aesthetic scores of pictures that the users have taken or selected. We designed and conducted a user study employing both quantitative and qualitative methods with AI/ML experts, photographers, and the general public. Through the study, we identified that (1) users understood the AI using their own group-specific expertise, (2) users reduced the thought gap with the AI by interacting with it through various strategies, and (3) the difference between users and the AI had a negative effect on interpretability and reasonability. Finally, based on these findings, we suggested design implications for user interfaces where AI algorithms can provide users with subjective information. We discussed the importance of synthesizing various perspectives in AI algorithms and interface design processes, as well as the possibility of exploiting various strategies and the need for mutual communication that users have shown when interacting with AI for both pedagogical purposes and to produce high-quality data. We hope that this work will serve as a step toward a more productive and inclusive understanding of users in relation to AI interfaces and algorithm design.

ACKNOWLEDGMENTS

We want to acknowledge and thank everyone who was involved in each stage of the research. We especially want to express our sincere gratitude to the anonymous reviewers.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 582.

[2] Christopher Ahlberg and Ben Shneiderman. 1994. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 313–317.

[3] Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016).

[4] JE Allen, Curry I Guinn, and E Horvtz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14, 5 (1999), 14–23.

[5] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349* (2015).

[6] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and information technology* 15, 3 (2013), 209–227.

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[8] Davide Castelvecchi. 2016. Can we open the black box of AI? *Nature News* 538, 7623 (2016), 20.

[9] Keunwoo Choi, George Fazekas, and Mark Sandler. 2016. Text-based LSTM networks for automatic music composition. *arXiv preprint arXiv:1604.05358* (2016).

[10] Dan CireşAn, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. 2012. Multi-column deep neural network for traffic sign classification. *Neural networks* 32 (2012), 333–338.

[11] Brenda Dervin. 2006. Project overview: Sense-Making Methodology as dialogic approach to communicating for research and practice. *Sense-making the information confluence: The whys and hows of college and university user satisficing of information needs. Phase I: Project overview, the Three-Field Dialogue Project, and state-of-the-art reviews* (2006).

[12] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. Ux design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 278–288.

[13] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 263–274.

[14] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.

[15] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; things can be worse than they appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Eleventh International AAAI Conference on Web and Social Media*.

[16] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 494.

[17] Leah Findlater and Joanna McGrenere. 2004. A comparison of static, adaptive, and adaptable menus. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 89–96.

[18] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 329–338.

[19] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *arXiv preprint arXiv:1806.00069* (2018).

[20] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 6645–6649.

[21] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623* (2015).

[22] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[24] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[25] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and others. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* 29, 6 (2012), 82–97.

[26] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 600.

[27] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 159–166.

[28] Eric Horvitz, Jack Breese, David Heckerman, David Hovel, and Koos Rommelse. 1998. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 256–265.

[29] Allen Huang and Raymond Wu. 2016. Deep learning for music. *arXiv preprint arXiv:1606.04930* (2016).

[30] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and others. 2018. Interpretability beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning*. 2673–2682.

[31] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems* 21, 5 (2006), 88–92.

[32] Will Knight. 2017. *The Dark Secret at the Heart of AI*. Technical Report.

[33] Carol C Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science* 42, 5 (1991), 361–371.

[34] LabintheWild. 2018. Augury Standalone. (2018). Retrieved September 17, 2018 from `http://augurydesign.com`.

[35] Vivian Lai and Chenhao Tan. 2018. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. *arXiv preprint arXiv:1811.07901* (2018).

[36] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science* 350, 6266 (2015), 1332–1338.

[37] Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A Terry, and Krzysztof Z Gajos. 2016. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4098–4110.

[38] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1603–1612.

[39] Dian Lei, Xiaoxiao Chen, and Jianfei Zhao. 2018. Opening the black box of deep learning. *arXiv preprint arXiv:1805.08355* (2018).

[40] Henry Lieberman and others. 1995. Letizia: An agent that assists web browsing. *IJCAI (1)* 1995 (1995), 924–929.

[41] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116, 1 (1994), 75.

[42] Richard E Mayer. 1996. Learning strategies for making sense out of expository text: The SOI model for guiding three cognitive processes in knowledge construction. *Educational psychology review* 8, 4 (1996), 357–371.

[43] Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and others. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 3 (2015), 530–539.

[44] Eric M Meyers, Karen E Fisher, and Elizabeth Marcoux. 2009. Making sense of an information world: The everyday-life information behavior of preteens. *The Library Quarterly* 79, 3 (2009), 301–341.

[45] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. ACM, 279–288.

[46] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2408–2415.

[47] C Naumer, K Fisher, and Brenda Dervin. 2008. Sense-Making: a methodological perspective. In *Sensemaking Workshop, CHI'08*.

[48] Changhoon Oh, Taeyoung Lee, Yoojung Kim, SoHyun Park, Bongwon Suh, and others. 2017. Us vs. Them: Understanding Artificial Intelligence Technophobia over the Google DeepMind Challenge Match. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2523–2534.

[49] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 649.

[50] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6620–6631.

[51] Google People+AI Research (PAIR). 2018. What-If Tool. (2018). Retrieved September 17, 2018 from `https://pair-code.github.io/what-if-tool/index.html`.

[52] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and others. 2013. Color image database TID2013: Peculiarities and preliminary results. In *european workshop on visual information processing (EUVIP)*. IEEE, 106–111.

[53] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184* (2018).

[54] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 103.

[55] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, 269–276.

[56] Reijo Savolainen. 2006. Information use as gap-bridging: The viewpoint of sense-making methodology. *Journal of the American Society for Information Science and Technology* 57, 8 (2006), 1116–1125.

[57] Chirag Shah. 2010. Coagmento-a collaborative information seeking, synthesis and sense-making framework. *Integrated demo at CSCW* 2010 (2010).

[58] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.

[59] Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810* (2017).

[60] Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*. ACM, 830–831.

[61] Glenn W Smith and Frederic Fol Leymarie. 2017. The Machine as Artist: An Introduction. In *Arts*, Vol. 6. Multidisciplinary Digital Publishing Institute, 5.

[62] John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization* 7, 2 (2008), 118–132.

[63] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[64] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[65] Hossein Talebi and Peyman Milanfar. 2018. Nima: Neural image assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.

[66] Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *Information Theory Workshop (ITW), 2015 IEEE*. IEEE, 1–5.

[67] MW Van Someren, YF Barnard, and JAC Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive. (1994).

[68] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 440.

[69] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 656.

[70] Optimal Workshop. 2016. Reframer. (2016). Retrieved September 17, 2018 from `https://www.optimalworkshop.com/reframer`.