Human-Centered Evaluation of Aesthetic Quality Assessment Models Using a Smartphone Camera Application

Samuel Goree
sgoree@stonehill.edu
Stonehill College
Easton, Massachusetts, USA

Jackson Domingo
jdomingo@students.stonehill.edu
Stonehill College
Easton, Massachusetts, USA

David Crandall
djcran@indiana.edu
Indiana University
Bloomington, Indiana, USA

Abstract

Human-centered artificial intelligence is caught between two approaches to subjectivity in evaluation, which can inhibit communication between engineers and designers on subjective problems. To better understand this tension, we study a test problem from computer vision: aesthetic quality assessment (AQA). We propose an approach to human-centered evaluation for this problem based on feminist epistemology, which transforms the benchmarking process into a design, engineering and user testing process. We demonstrate this approach for AQA by designing a smartphone camera application that takes photos based on the output of an AQA model. Through a user study, we examine both the performance of our interface and the underlying models. We find that a design goal of legibility is crucial for the success of both the interface and the underlying models, and recommend that human-centered evaluations are integrated early into the modeling process for these problems, before formalizing the problem statement.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; **HCI theory, concepts and models**; **User centered design**; • **Computing methodologies** → *Machine learning*; *Computer vision.*

Keywords

Aesthetics, feminism, situated knowledge, evaluation, AI, smartphone, camera, legibility

ACM Reference Format:

Samuel Goree, Jackson Domingo, and David Crandall. 2025. Human-Centered Evaluation of Aesthetic Quality Assessment Models Using a Smartphone Camera Application. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3715275.3732209

1 Introduction

Today, there is a pervasive tension between artificial intelligence (AI) and human-computer interaction (HCI) over subjectivity that complicates evaluation in human-centered artificial intelligence (HCAI) [55]. Specifically, machine learning (ML) practitioners tend to view subjectivity as external to evaluation while HCI researchers tend to view subjectivity as central to user experience [39] and thus internal to evaluation.

This tension is clearest during the handoff from ML engineering to user experience (UX) design. Consider the following hypothetical scenario: Alice is a machine learning engineer. She is developing a new user-facing ML model for her organization's product. She re-implements several models proposed in the literature, training each on historical user data and measures their quality using labels derived from that data. She finds one model to be the most accurate, and hands it off to Brenda, a front-end designer and UX researcher. At the same time, Brenda has been prototyping a front-end interface, and once she has Alice's model, she conducts a user study. But, to her surprise, she finds that the new feature confuses users. Brenda thinks the model behavior is to blame, but Alice insists it is objectively the best model.

Ultimately, this tension is due to a difference in evaluation standards with respect to subjectivity. Alice has defined the problem statement and training data based on a specific operationalization of the intended behavior, which Brenda observes is inconsistent with users' expectations. Clearly, the issue is a lack of coordination between Alice's model development process, Brenda's design process and users' expectations. But how should they bring their mental models into alignment? To investigate this issue, we study a highly subjective test problem: image aesthetic quality assessment (AQA). AQA seeks to apply machine learning to measure the aesthetic quality of photographs and other images, usually by classifying them as "high" or "low" quality. The models are typically trained and evaluated based on the opinions of human raters, for example, from a photography challenge website [11, 40, 56] or a crowd work platform [44, 62]. While aesthetics may seem solidly outside the domain of computation, there are a number of reasons researchers would like to have aesthetic quality measures, including for direct application in automatic photo curation and editing [49], as well as for indirect use in evaluating image generative models [72] and image processing algorithms like computational bokeh effects [32]. When AQA models are used for evaluation or data curation behind the scenes of another process, they become infrastructural [68] and invisible. These models are also interesting for scholars of AI art, as the LAION Aesthetics v2 dataset[1] is curated based on AQA research.

Researchers currently evaluate AQA models using performance benchmarks, such as the analysis of visual aesthetics (AVA) benchmark [56]. AVA consists of images and aesthetics ratings from 1–10,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
FAccT '25, Athens, Greece
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1482-5/2025/06
https://doi.org/10.1145/3715275.3732209

[1]This is an open dataset used to train many artistic image generation models. For more information see here: https://laion.ai/blog/laion-aesthetics/

which are averaged and thresholded to arrive at a binary class label (high or low). Some users may find that the ground truth aligns well with their sense of taste, but others may find that it disagrees considerably — up to 30% of benchmark rankings may be inconsistent with any given user's taste [26]. Human judgments are also unstable and evolve as our taste changes at all stages of life [61], so it is not guaranteed that benchmark labels will be accurate for the same participant in the future. We believe a new evaluation paradigm with a theoretically grounded concept of subjectivity is necessary to bridge the gap between UX and ML for these subjective problems.

We propose *situated evaluation*: a method for human-centered evaluation in AI grounded in Donna Haraway's posthumanist and feminist concept of situated knowledges [30]. We demonstrate a situated evaluation of four aesthetic quality assessment models and present our qualitative results. Finally, we analyze our evaluation process using the concept of legibility and generate recommendations for future situated evaluations.

1.1 Defining Situated Evaluation

Given a prediction problem with fixed input and output data and several competing methods (e.g. model architectures or training datasets) for solving that problem, we propose carrying out the following three steps:

(1) **Design:** Design a user interface which allows the user to collect input data in real time and see the output of one or more AI methods.
(2) **Engineering:** Implement the design along with several competing methods for this problem so that their output can be viewed in real time alongside changing input.
(3) **UX:** Conduct task-based user studies of this interface where human participants are tasked with identifying the advantages and disadvantages of each competing method.

We consider all three of these steps as knowledge-generating processes. The first stage allows the designer to generate knowledge about the design space created by their problem statement, and can help them find limitations or hidden assumptions that prove problematic for designers. The second stage helps the engineer to understand the implementation and performance requirements needed for their AI methods to be production-ready. The third stage generates qualitative data comparing ML methods and integrates non-expert perspectives into the design and development process early, in the tradition of participatory design [54, 65].

The end goal of this process is not to develop a market-ready product, as would be expected in UX (though if the design in step 1 is particularly good, a real product design may come from it). Instead, it is to evaluate the "user-readiness" of the underlying technology in qualitative terms and to provide an avenue for communication between UX researchers and ML practitioners about performance while ML model development is still underway. Importantly, we find that this method also provides an avenue for designers, engineers, researchers and users to push back on the assumptions behind the problem statement, a form of evaluation which is not typically possible in ML.

2 Related Work

2.1 Evaluating Aesthetic Quality Assessment

Aesthetic quality assessment has been studied in computer vision for almost two decades. The early work in AQA of Datta et al. [11] and Ke et al. [40] used relatively small datasets of images with binary quality labels and measured performance in terms of classification accuracy. In 2012, Murray et al. [56] released the Analysis of Visual Aesthetics (AVA) dataset, which contains over 250,000 photos from DPChallenge.com along with metadata such as rating distributions and category labels. This dataset continues the original framing of AQA as a classification problem, where the ground truth label (high or low quality) is calculated from the average of many human ratings. Recently, others have framed AQA as a more subjective problem. Ren et al. [62] introduce the personalized image aesthetics task through the Flickr-AES dataset, which contains user-by-user ratings for each image.

Progress on aesthetic quality assessment benchmarks suggests that recent models more accurately classify between high and low quality images, regardless of individual subjective differences between users. However, prior work argues that labels from these datasets fail to capture the concept of aesthetic quality broadly, and instead captures a *specific* "aesthetic" photographic style common on photo-sharing websites [25]. Arguably, this style has become the default style for commercial AI art generators. The relevance of this "aesthetic" photo style to individual aesthetic preferences varies considerably from user to user, and personalization only increases that variance [26]. These issues cast doubt on the reliability of benchmark-based evaluation for this problem.

Algorithmic analysis of art and aesthetics has been the subject of criticism from the humanities. Recently, Ramya Srinivasan identifies tensions between scholars of the visual arts and AI researchers surrounding algorithmic curation and generation of art, specifically related to stylistic reproduction and the formation of a canon [67]. Michelle Elam identifies that in its approach to the arts, AI reproduces limiting models of the human based on racial pseudoscience, limiting achievement, expression and progress. She argues that AI research needs deeper integration with humanistic approaches to difference in order to contribute to human flourishing [20].

Similarly, benchmark-based evaluation has been the subject of criticism from science and technology studies. Recently, Orr and Kang present a genealogy of benchmarking, linking it to competitive sport. They find that the subjectivity of new AI tasks and lack of measurement for factors like societal impact, ethical concerns, and practical applicability will require us to reevaluate whether benchmark results are meaningful [59]. Similarly, Denton et al. argue that benchmarks are inseparable from their annotators, and call for increased study of annotators in machine learning [13–15].

2.2 Interpretable AI

Approaches to explanation and visualization for computer vision models typically focus on offline approaches. These include visualizations of feature maps [71], class activation maps derived from deep neural networks [66], nonlinear dimensionality reduction techniques like T-SNE [70] and UMAP [53] for visualizing feature spaces and more sophisticated agent-interaction explanations like those of Hendricks et al. [35]. Outside of computer vision, there is a

robust literature in visual analytics for interactive visualization for machine learning model development; see [63] for a literature review. The goals of visualization depend on user psychology; model explanations should cohere well with how users already mentally model the problem under study [50], and there is often a tradeoff between soundness and completeness in explanations which shapes user understanding [47]. We approach this problem more like the interactive visualizations, allowing the user to build their own intuitions and explanations based on interaction with the model.

2.3 Design Research for Photography

We employ concepts of research through design [22], speculative design [18] and critical design [5, 7]. These related methods use design practice as a form of research, speculating about alternative futures and explicating the implicit assumptions behind technology. Several papers in this area explore the consequences of different camera designs. For example, Odom et al. apply these methods in their design of Photobox, a speculative slow technology which occasionally prints photos from a Flickr library, slowly amassing a collection over years, which challenges the notion that technology should be fast, easy and disposable [58]. Pierce and Paulos's Inaccessible digital camera, a camera made of concrete which must be destroyed to extract the digital storage medium within, similarly questions notions of functionality and disposability [60]. Recently, Karmann's Paragraphica continues in the legacy of speculative camera designs. This camera-like device lacks a lens or photo sensor, instead creating photographs using an image generation model conditioned on location, time, temperature and maps data, questioning the mapping between photos and reality [38]. The important characteristic is that these designs are not potential future products, but instead are used to explore alternative design spaces and make abstract critiques of technology tangible, facilitating future designs in these spaces.

2.4 Feminist Epistemology

AQA is difficult because the aesthetic quality of images is subjective. We advocate a concept of subjective knowledge for this problem guided by Donna Haraway's concept of situated knowledges [30]. Within her larger post-humanist feminist project, Haraway deconstructs boundaries between humans, animals and machines [29]. A consequence of breaking those boundaries is blurring between objective and subjective knowledge. For Haraway, all knowledge comes from one view of the world, situated in time and space, mediated by biological, mechanical, social, cultural and political factors. Science, when it separates a view of the world from the way in which it was captured, performs a "god trick" — pretending that an observer's limited view can actually see everything from an omniscient god's-eye view. Feminist objectivity acknowledges each observer's perspective, and does not elevate supposedly objective knowledge over other forms of knowledge when finding the truth.

Under this concept of subjectivity, ML methods for AQA are computational ways of seeing images, rather than more objective formulations of image aesthetics or artificial subjects experiencing aesthetic phenomena themselves. In the context of ML, researchers who treat evaluation benchmarks as objective measures of model performance are performing a god trick as well, and separate our knowledge of performance from the humans who generated the benchmark data.

Situated approaches to knowledge were introduced to HCI by Bardzell and Bardzell [6], and have led to a wide variety of qualitative design studies which center the experiences of women and prioritize situated and embodied approaches to knowledge. Similarly, situated and embodied approaches to knowledge are a pillar of D'Ignazio and Klein's data feminism [16, 43]. Several recent papers have critiqued god tricks in feature importance metrics [28], knowledge-enhanced language models [45] and image classification benchmarks [13].

To operationalize Haraway's theory, we turn to secondary literature. Bhavnani, writing in the context of women's studies [8], proposes three criteria for feminist objectivity:

(1) Reinscription: Does the research method portray the participants as passive and powerless, or does it recast them as active agents?
(2) Micropolitics: Does the research engage with the political relationships between researcher and participant?
(3) Difference: Does the research engage with differences in perspective between participants?

Our approach to evaluation takes place in the world at the time the photograph is taken. There is no hidden photographer responsible for the images. Evaluating at the time of photography avoids the first god trick because we do not disconnect the very literal view of the world from its source. On the more metaphorical level, we also meet Bhavnani's criteria:

(1) We recast the human subjects, who in other AQA research are anonymous crowd workers, into active participants in the research process who are given space to express their nuanced views about aesthetics and cameras.
(2) By giving participants evaluative agency, we reverse the typical power dynamic in machine learning where photographers and labelers are disconnected from the models derived from their data. Instead of being judged by AI, our participants are judging it.
(3) Qualitative analysis gives us the flexibility to handle subjective difference with nuance, rather than reducing it to a measure of dispersion or personalization technique.

3 Methods

Based on our concept of situated evaluation, our study consisted of three phases: design, engineering and user experience (UX) evaluation. In this section, we describe the methods used in each of these phases.

3.1 Design

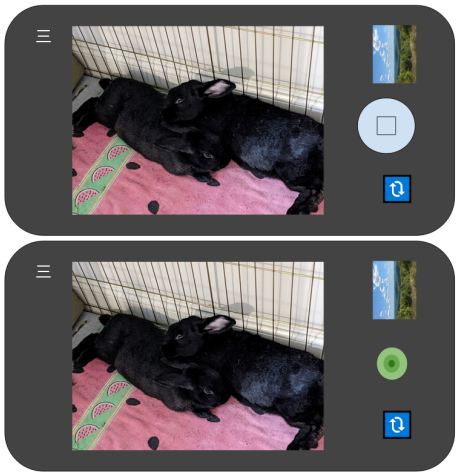
Inspired by Haraway's concept of knowledge situated in a specific time and place, we pursued a design goal of making aesthetic quality assessment algorithms tangible, so that participants can judge the algorithms' learned aesthetic preferences in a real-world setting. A low-fidelity prototype, an image we showed on a smartphone to test the concept internally, is shown in Figure 1 (a).

The application was developed on top of the existing open source project Open Camera, licensed via the GNU GPL [31]. We added three settings to the application:

(1) "Aesthetics Capture Mode" removes the shutter button and starts a background process that takes and evaluates one photo per second. If the estimated IAQA score is greater than a threshold, the photo is saved and a visual feedback animation for photo capture plays.
(2) "Aesthetics Indicator Mode" starts the same background process, but adds a line plot to the top of the preview showing the IAQA model output. If aesthetics capture mode is also on, the plot shows a horizontal threshold line as well.
(3) "Aesthetics AI Sensor" allows the user to choose one of four models, described below.

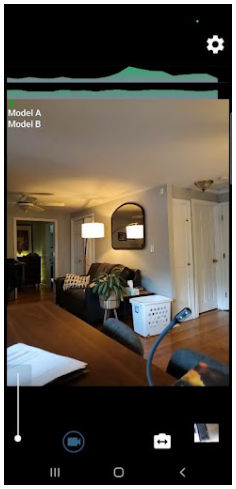
Removing the shutter button was a key decision rooted in critical design methods. Users assume they have full control of a camera, but many of the choices involved in photography are made automatically in cameras already [36], simulating different film ISO values [27] and creating computational Bokeh effects without adjusting focal length [32]. These auto settings are often adjusted based on computational image quality measures [32]. Users, however, expect to choose when photos are taken, and removing that aspect of user control could lead them to reconsider the way in which computational measures of taste are already influencing photography, and to speculate about human-AI co-creativity in photography [12] and possible future AI art forms. More practically, giving control of the camera to an AQA model makes the model unavoidable, and makes obvious its (in)consistency with the users' preferences.

We experimented early in the development process with removing the camera preview and using different kinds of visual or haptic feedback, but found in initial testing that taking photos without a preview was difficult. We also moved away from the visual indicator present in the early design (Figure 1 (a)), which looked too much like a shutter button, leading our pre-pilot user to tap it and expect to take a photo. Haptic feedback proved especially unusable because it was difficult to calibrate: either the phone would vibrate constantly, annoying the user and draining the battery, or vibrate too little to be helpful for interpreting the model output. As a result, we designed the minimal line plot shown in Figure 1 (b). After conducting five user studies, we iterated on our design, simplifying the interface and adding a second line plot and second sensor option to the menu to make it easier to compare two models. Our final design is shown in Figure 1 (c).



(a) Low fidelity prototype: a smartphone camera app with and without a shutter button.

(b) Screenshot of the first prototype.



(c) Screenshot of the final design with two models.

Figure 1: Three design iterations for our application.

3.2 Engineering

Our prototype implements four specific models which are characteristic examples of four different approaches to aesthetic quality assessment, inspired by different eras of research on this problem. While these models are inspired by particular previous works, we implemented each model ourselves from scratch and do not claim to be evaluating the work of other specific authors.

A: A baseline model using the mean of the output of an approximate image Laplacian filter applied to the image. The output of a Laplacian is highest in areas where there are sharp visual edges, and decreases in areas that are blurry, giving it a slight positive correlation with aesthetic quality.
B: A linear model using hand-crafted features based on the early AQA work of Ke et al. [40]. This model uses four sets of image transformations: the image's Laplacian, 4096-bin color histogram, Fourier transform and lightness distribution. The first two transformations are further distilled by taking the mean feature map of the positive and negative classes and measuring the $\mathcal{L}_1$ distance from the test image to the mean for each class. In the case of the Fourier transform, we follow Ke et al. and measure the highest frequency bin with value greater than 5. For the lightness, we measure the width of the 98% mass distribution. While Ke et al. use a Naive Bayes classifier on these features, we use logistic regression for ease of deployment.
C: A 2014-era deep neural network, based on the 8-layer AlexNet architecture [46], with a two-column approach similar to Lu et al. [52]. To avoid warping images to the 224 by 224 resolution required by the AlexNet architecture, this model has one network for a center cropped local view at full resolution

and another center cropped global view downsized to 256 by 256. Following Lu et al., we concatenate the models' hidden representations before applying the final fully connected layer.

D: A more contemporary deep neural network approach, based on the 18-layer Resnet architecture [34] and trained using the Adam optimizer [41], without any other AQA-specific modifications.

Both deep neural networks are randomly initialized using He initialization [33] and trained on the AVA dataset using a cross-entropy loss function and learning rate starting from 0.001 and decaying multiplicatively each epoch by $1-10^{-7}$. We emphasize that our goal here is not to qualitatively test specific modeling decisions. Instead, we use these four models as examples which characterize four different approaches to the problem. To take photos, we use an adaptive threshold, so a photo is taken every second, but it is only saved if it is rated at least 10% higher than the average of the prior 10 photos.

Quantitative evaluation results for the models using the AVA dataset benchmark are shown in Table 2. Following the AQA literature discussed in Section 2.1, we report accuracy, ranking correlation (Spearman's ρ) and ROC-AUC (area under the Receiver Operating Characteristics curve). All metrics are computed on the AVA test set. Despite the fact that the model architecture was not designed for the task, Model D performs the best across the board. While our first three approaches roughly mirror the reported accuracy in their respective papers, none of these models are as accurate as more recent state-of-the-art methods: we encourage future work evaluating specific contemporary approaches.

3.3 UX

After developing our camera interface and receiving IRB approval, we conducted a user study of situated evaluation for AQA. We conducted one 40-70 minute session with each of twelve participants. Five participants had their sessions between February and March 2023. At this point, we revised our camera interface based on their feedback and conducted seven additional sessions in October 2024 to reach inductive thematic saturation [64]. Participants were recruited using paper fliers and university mailing lists, and paid US$15 for participation. All participants were graduate or undergraduate students with no expertise in computer vision, AI or machine learning; see Table 1 for details. We focused our initial recruitment on students with some graduate training in HCI, but relaxed that focus after session 4.

Sessions were conducted in public spaces on three American college campuses, one in the midwest and two in the northeast. Each session was conducted according to a semi-structured protocol in three stages:

(1) The facilitator briefly describes the premise for the study and guides the user through the interface.
(2) Once the participant is comfortable using the interface, the participant is asked to wander around the space and take photos using each of the four models. While we did not tell the users which photos to take, we suggested photos of scenes, small objects, large objects, bright colors and people.
(3) Once the participant affirms they have a good understanding of how the models are similar or different from one another, the participant and facilitator sit down and review the saved photos. Finally, the facilitator asks brief closing interview questions regarding similarities and differences between the models, specific peculiarities of each model and usability of the app.

The final conversation of each session was audio recorded and transcribed. We also saved photographs from each of the sessions. Our analysis followed a constructivist grounded theory approach [10, 24] through inductive content analysis [21, 48]. Our goal in analysis is to develop a theoretical understanding of how participants interact with AQA models, and establish methodological recommendations for future, similar studies which seek to evaluate these models. Grounded theory is an appropriate methodology for this task because it prioritizes theory generation over theory confirmation. Additionally, we collect multimodal data, including images and transcripts, which content analysis is well suited to approach. Practically, we took notes on our transcripts using a word processor comment feature, cross-referencing images based on timestamps. After identifying several cross-session themes in open coding of sessions 1–5, we engaged in a second analysis of the transcripts and images to locate characteristic examples of each theme, with an emphasis on adjectives that participants used to describe models. After the second round of sessions, we analyzed sessions 6–12 similarly, then engaged in a third round of coding to identify common themes across all sessions.

4 Findings

In this section, we describe the results of our user study sessions, integrating findings from the design and engineering phases as appropriate. We start by addressing participant perceptions of the different models' performance, then go through four additional themes.

4.1 Perceptions of Performance

Our participants tended to judge the models' performance based on consistency with the photos they would or would not have taken. We categorize these inconsistencies into false positives and false negatives.

In terms of false positives, P2 was annoyed that model A took *"more photos than I would have taken."* P6 pointed out that models A and D were *"too sensitive"* while models B and C were *"too selective"* and P1, P7 and P12 all described at least one model as taking *"random photos."*

In terms of false negatives, participants found it extremely frustrating when they wanted to take a photo but the model would not cooperate. P6, P8 and P9 initially mistook the lack of photos being taken for *"lag"* (P6) or *"buffering"* (P9). P1, P9, P10 and P11 all noticed that models A and B would not take a photo of the interlocutor. These models are based on simple image features based on edge and color distributions and do not take human detection into account. P6 and P11 remarked that there was a photo they wanted to take but *"the model just wasn't seeing it"* (P6). P10 said he was not sure what the models were thinking and wasn't sure that models A or B were working.

Table 1: Study Participant Demographics. We specifically recruited students at three universities without expertise in computer vision.

Participant	Age Range	Gender	Academic Background
P1	31-40	Woman	HCI, health informatics, computer science
P2	31-40	Man	Security informatics, HCI
P3	31-40	Woman	HCI, data science
P4	21-30	Man	HCI, computer science
P5	31-40	Man	Electrical & computer engineering
P6	21-30	Woman	Philosophy, criminology
P7	21-30	Man	Psychology
P8	18-20	Man	Marketing, data analytics
P9	21-30	Man	Sports management
P10	21-30	Man	Communication
P11	21-30	Man	Political science, psychology
P12	21-30	Man	Business management, Spanish

Participants were divided as to whether false positives or false negatives were more problematic. P1, P2, P8, P10 and P11 all would prefer more false positives. P1 *"would rather want to have pictures there in my hand for me to sort and like delete the ones that are not good."* P8 would prefer that it took photos immediately upon opening the app rather than waiting for a good shot. On the other hand, P3, P4, P6, P7 and P9 would prefer that the app take fewer photos. P3 does not ever want to take more than one photo: *"I just take one photo and send. I don't take even two and choose one because I don't need that."* P7, using contemporary slang, observed that any model taking too many photos is *"kind of cooked;"* a bad thing. P9 became upset when he realized that all the photos the model was taking would end up in the phone's camera roll. He wants to retain control of which photos are actually saved alongside the photos he has taken and doesn't want an automated system to access his photo library without a manual "filtering" step. P12 applied an information retrieval framework to his description of model performance: *"when you're at a sporting event or whatever, I guess it is more convenient than taking photos, but at the same time, you don't want to inconvenience users by showing them photos that are completely irrelevant. Kind of like an AI. Like when you give it a search prompt. How you don't want it to show search results that are completely irrelevant to what you're asking."* For him, the model only provides a convenience if it selects "relevant" photos.

4.2 Personification

We found that half of our participants (P1, P2, P4, P6, P7, P10, P11) had a tendency to personify the models. False negatives were often described using language around the model not "seeing" their vision for a photograph. P4 observes that *"[model] B doesn't seem to be very excited, it's almost like, very stoic."* P1 constantly turned to language around the models' likes and dislikes (*"C did not like this but A liked it...It definitely loves patterns. Like uniform patterns."*) and interests (*"Now I cannot be too sure if it is [taking photos] because of the floor or if it is because of the chair because it was finding the floor very interesting."*). While the way in which computers can function as social actors is well-discussed in HCI [19, 23, 57] and AI [19, 69], we emphasize that this personification is happening without the use of natural language or a human-like artificial persona. Just a letter name and a scalar measure of "preference" was enough to lead these participants towards these patterns.

To illustrate the differences in "personality" our participants observed, we collected all of the words used to describe each model's character or emotions. Descriptive words are shown in the right column of Table 2. We can see that participants use terms like *"picky"* or *"stoic"* to describe model B's low variability, and terms like *"difficult to predict," "random"* or *"like a cop camera"* to describe the high variability and preference for red cars of model D.

4.3 Interpretability

Several of our participants had a tendency to try to interpret what the different models were doing to arrive at their judgments, conceptually reverse-engineering the models. For example, P5 offers a number of ideas: *"is that it? How close it is to the people?" "it's taking pictures of trees and...still objects,"* and *"does it also...try to understand the color?"* P3 speculates on the content of the training data: *"is it just trained on like landscapes and not people?"* P2 identifies that model B seems to be obeying photography rules: *"[model B] tries to actually pick larger objects that are sort of center focus, which would be like focal points...I have no idea if this uses like the rule of threes for how you frame up stuff or not."* Interestingly, Model B, based on Ke et al. [40], is explicitly designed based on these photographic rules of thumb. P2 also notes that model D prefers specific colors of cars: *"apparently it really likes blue or red vehicles. Which makes me think that's probably because, it's probably trained on those the most."*

This reverse-engineering approach led to feelings of confusion and disappointment when models C and D were revealed to be deep neural networks, not interpretable visual measures. While all our participants were aware of AI, it is unclear how familiar each one was with the limitations of different modeling approaches. P3 in particular dislikes that the models are not clearly nameable: *"Instead of model ABCD, how about you write something like model nature, model human, model bird?...Ok you give me three words, you know, what is model D's feature?"* P3 wants to evaluate the models in terms of what they actually measure, since she does not think that

a single, general kind of aesthetic quality exists, and is frustrated that such descriptions are not possible for the deep neural network models.

Interestingly, this lack of interpretability led both P4 and P8 to express a desire to trust the model's taste over their own. P4: *"assuming that machine learning models know much more than us, even though they might not think like us, but they have a larger set of data that's trying to feed them...I would want it to be more selective."* P8 wanted AI-based photography advice: *"I think having this to kind of help to decide this angle is good. This lighting is good, that sort of thing."* These comments echo findings regarding a novel trend towards implicit trust of algorithmic systems as ultimate authorities [37].

When asked to choose a favorite model, a majority of participants preferred either model C or D, one of the deep neural networks, over models A and B, often identifying either C or D as their favorite. Their justifications centered around their *"consistent"* and *"picky"* behavior, taking fewer photos than other models. See Table 3 for details. Interestingly, preferences for model C contradict benchmark scores, which have it performing significantly worse than model D. Benchmark scores and participant adjective descriptions are shown in Table 2.

4.4 Suggested Use Cases

As expected for a critical design project [5, 7, 18, 22], participants were frustrated by the AQA models and frequently suggested better use cases for the technology. For example, P8 and P10 suggested that users would prefer a system that offers photography advice above a system that takes photos automatically. P1 and P5 suggested that users could use this technology to take photos more easily when their hands are busy, such as while driving. P9 pointed out that this technology could be used to take group photos without using a timer. P12 suggested designing for press photographers who need to take specific photos while their attention is elsewhere. These behaviors echo the concept of participatory requirements specification [54].

4.5 Perspectives on Photography and AI

Two participants referenced existing photo editing applications during the study to explain their perspectives. P3 references Meitu, an app described on the Google Play store as "Make your photos stunning and sensational! Whatever your beauty preference, do it all with Meitu!" [1]. P3 elaborates, *"You don't need to do any makeup, it's makeup for you! So a lot of girls including me like this because sometimes we don't need to make up, but we can make up here, you know. It makes me white, and it makes me clear, and it removed the dark part of my face or the environment or this is like, makes me younger."* The main appeal of this tool, for P3, is that it has a huge variety of filters and editor features so that each user can find the combination which looks best to them. She would not use any kind of photo tool unless it gave her that kind of aesthetic control.

P4 references several other applications, including the social media platform Instagram and VSCO, an app described on Google Play as "a leading photo and video editor that nurtures the creative journey with our library of 200+ premium quality presets and tools" [2]. P4 describes how he would edit one of the photos taken by our interface, *"this also could be considered aesthetic, like if I was trying to post this to Instagram, I'd like blow out the highlights and make the background look a bit more even."* Editing is core to his photographic practice: *"I look at all of them as starting points and what can I make that goes beyond what I took."* This approach shapes the way he looks at the photos taken in this study: *"I don't understand some of the reasons for these shots. Like it could be made aesthetic...by aesthetic I mean things that could like possibly go on to Instagram."*

Participants also had a variety of nuanced criticism of aesthetic quality assessment. P4 believes that the way computers and humans judge photos should remain complementary:

> *"I find it harder to figure out if a machine can think of aesthetics in the same way that humans do because, for a machine, [the photo] is the final picture, but for human that is not the final picture and we can always like step it up and make it look more interesting. So if it's a sort of discerning person, like probably a designer or a photographer, and they might just like be inspired...When working with AI...you work in conjunction, one doesn't replace the other and basically things that might take up a lot of time or like, instead of grunt work that you can leave to the AI and then you can use it as a sounding board or like an inspiration to get to something that's more refined and polished."*

In other words, even if an AQA algorithm is used to evaluate photos, the final say regarding aesthetic quality should remain in the hands of the user, which echoes P9's sentiment about the camera roll.

Similarly, P3 and P11 were extremely critical of aesthetic quality assessment. P11 expresses his perspective:

> *"I don't really care what an AI thinks is beautiful or what an AI thinks is nice. I care more about what I think is nice or beautiful or good, you know?...I don't know if it's right for AI to be deciding what's aesthetic and what's not. Like morally speaking...specific tastes get washed away in light of some objective. You know, if you have AI you can say...this is the most aesthetically pleasing painting of all time. Or you know what I mean? Like, I feel like reducing things down to the numbers and like statistics sort of limits the inherent value of anything that's subjective...when it comes to things that are aesthetic, there's something greater than what could be measured in numbers. That there's something deeply human, and I would even say, like, spiritual, about art about things that are beautiful as a whole that I think is really hard to kind of measure out."*

For him, human intention is essential for photography. The reasons that people take photos are personal and unquantifiable, due to the spiritual nature of aesthetics. He does not care what an AI thinks is beautiful. P3 concurs:

> *So that shows the model's emotion? Then I need to satisfy the model not satisfy me. Yeah, this model used me, not I used the model...I need to understand the algorithm behind more like what makes this number peak? What brings the value down? If I don't know the calculation, I just don't understand what...[if] the algorithm of the*

Table 2: Accuracy, Ranking Correlation and AUC metrics for each model on the AVA test set, juxtaposed with the adjective descriptions used by our participants for each model.

Model	Accuracy	ρ	AUC	Descriptions
A	0.600	0.047	0.530	*very selective* [P1], *greedy* [P2], *strange* [P4], *unreliable* [P4], *sensitive* [P6], *reactive* [P7], *conflicted* [P10]
B	0.703	0.059	0.500	*low threshold* [P1], *picky* [P2,P4], *stoic* [P4], *unbothered* [P4] *jagged* [P6]
C	0.708	0.296	0.546	*loves patterns* [P1], *picky* [P4], *understandable* [P5], *smooth* [P6], *object recognition* [P4,P7], *responsive* [P8], *consistent* [P10]
D	**0.740**	**0.473**	**0.605**	*unpredictable* [P1], *random* [P1], *like a cop camera* [P2], *likes most things* [P4], *object oriented* [P5,P7], *responsive* [P8]

Table 3: Responses when asked to choose a best model. Participants who did not make a single choice are listed with a slash.

Participant	Choice	Reasons
P1	B / D	*"I don't want a model that is so selective...I want to have pictures for me to sort and delete...D is at least choosing human faces, B is not even doing that."*
P2	C	*"You're not picking up everything, but you're also not having such a low reaction rate that you don't pick up anything."*
P3	N/A	*"I don't mind to click the shutter button because that is the certain moment and the angle I want to take it! I definitely need that moment! I don't want the camera to take it for me."*
P4	C	*"What I would want from that model is to take an unexpectedly nice picture, which means I would want it to be more selective if it's going in conjunction with the manual button."*
P5	C	*"Model C is taking picture when there's some sort of like, nature...it's like a landscape photography...feature."*
P6	C / D	*"I thought it was just smoother, but it was a little harder to get a picture of what I wanted."*
P7	D	*"It felt like it actually was reacting to those objects in like a notable way...The reaction was kind of consistent."*
P8	C / D	*"The second [pair] were a little quicker with taking pictures. So I think that was a little more useful, helpful."*
P9	C / D	*"It was actually capturing the things I was trying to capture...The first [pair] never really wanted to take a photo of exactly what I wanted."*
P10	C / D	*"Took pictures more frequently, which was good...makes it feel like it's actually working.*
P11	N/A	*"A/B were definitely a lot faster to recognize things in like the natural world...C/D to me also seems like it was a little bit easier to work with things that were not necessarily natural...I didn't, while I was doing it, notice a gigantic difference between them."*
P12	A / B	*"A and B seem to like less pictures that were kind of irrelevant."*

> *model preference is not that good?...If the model itself is not good, I don't need to satisfy that model. Or maybe when, the moment I don't satisfy the model is the correct thing or is a good thing to do! You know, if the model is not the best one, there is no need to make it high."*

In other words, P3 gets to the heart of the concern shared by P4 regarding user control: if the camera has the final say on whether a photo is taken, it shifts the balance of power, placing the user's behavior under the algorithm's judgmental gaze: *"this model used me."*

5 Discussion

To summarize, we designed an interface for evaluating AQA models in the real world, and conducted a pilot study of its effectiveness. We designed our interface as a camera application with no shutter button and settings for four models. In our study, participants tended to evaluate the models based on false positive and false negative rates, but disagreed over which was more significant. Several participants tried to figure out how the models worked, either by personifying or reverse engineering them. Finally, several participants related their judgments back to their prior experience

with smartphone photo editing apps and expressed strong opinions about aesthetic quality assessment more generally. Our interface was successful at gathering participant feedback, both on the qualitative aspects of the models' performance as well as the potential issues related to the concept of AQA itself which are not revealed in benchmark-based evaluations.

5.1 Legibility

During our initial prototyping stage, we noticed that a variety of designs such as haptic feedback and a changing circular visual indicator were distracting. After our first round of user studies, we also found that users found additional camera settings to be distracting. We ended up making design choices prioritizing the *legibility* of the underlying model. We came to realize that the concept of legibility is useful for situated evaluation on multiple levels. As such, we synthesize two definitions: one from typography [9] and one from human-robot interaction research [17, 42].

First, we can use legibility to think about our camera application as less like an app and more like a medium. In typography, legibility refers to the factors that affect reading performance for sighted subjects [9]. The most legible typefaces are clear and do not distract from the text, but are still visible and contribute to the design. An interface for situated evaluation should work the same way: it should be easy to read, avoiding design elements that distract from or obscure the underlying AI model, but without trying to be invisible or neutral. (Lindley et al. reach a similar conclusion regarding AI visual imagery [51].)

Second, we can use legibility to think about the behavior of the underlying AI model. Alexandra Kirsch defines legibility in a robotics context based on two criteria: (1) The human observer or interactor is able to understand a robot's intentions and (2) the behavior meets the expectations of the human observer or interactor [42]. In other words, an AI system communicates intentions through behavior and behaves according to expectations. In our context, this provides a theory of evaluation for the AI model under study. The model is working if and only if the human evaluator is able to understand its sense of taste and its behavior met the evaluator's expectations for such a system.

These two concepts of legibility are coherent with Haraway's theory of situated knowledges. In a robot, camera app or typeface, a legible interface is a lens through which we are seeing a behavioral intent, sense of taste or text, respectively. While we can study legible typefaces scientifically and arrive at principles for legible design, there is ultimately no concept of legibility without a perceiving subject. And no matter how transparent the interface is, it will always be present mediating our interactions. Rather than creating transparent interfaces that fool the user into thinking they have full knowledge of the underlying model, we encourage legible interfaces which make users aware of the limits of their vision.

5.2 Participant Evaluations

Participants' evaluations were shaped by the implementation differences between the models, the context and protocol of the study, as well as their differing backgrounds and prior relationships with smartphone cameras. Models A and B, due to their implementations, tended to produce values with lower variance and less consistency with user preferences. Participants read that behavior as selective, picky or stoic. Model C, while not the highest performing quantitatively, was viewed as understandable, responsive and consistent. Model C was thus the most preferred model, chosen by 7/12 participants, followed closely by model D, chosen by 6/12. In other words, participants preferred model C because it was the most legible of the four models: they were able to understand its intentions and its behaviors were consistent with those intentions.

We found that participants' preferences were mediated by their prior experience with photography. This is best illustrated in the contrast between P1 and P2's responses to false positives. For P2, models *"took more photos than I would have taken."* While P1 was similarly frustrated by false positives, she preferred it because she *"would rather want to have pictures there in my hand for me to sort."* Similarly, P6 and P11 became frustrated when they wanted a photo to be taken but the model *"just wasn't seeing"* it. Finally, the design of the interface brought our work into comparison with other camera applications. P3 and P4 were frequent users of other smartphone photography tools, and interpreted our interface based on their prior experience of those tools.

Participant evaluations were also shaped by a tendency to personify the different models and ascribe mental and emotional states to them. This tendency is well-known as the ELIZA effect, after the 1964 text-based AI therapist, which human participants believed to have empathy even though it was only procedurally generating responses. Hamid Ekbia claims the ELIZA effect is an example of a broader "attribution fallacy" where humans believe that a computer system has mental faculties and emotional states much like their own, even when the system demonstrably does not [19, p. 8]. While the ELIZA effect has been observed in chat programs, case-based reasoning systems [19, Ch.5] and social robots like Kismet [69], it is surprising to see in a system as inhuman as a line plot above a camera window.

These differing interpretive factors are the heart of our concept of situated evaluation: participant evaluations are not fully determined by the objective characteristics of the models, but they are not fully subjective either. Instead, they are a product of the research environment, mediated through the model, interface and participants' way of seeing. In continuity with Haraway, we do not recommend attempting to eliminate these confounding factors. Instead, we recommend considering evaluations in context qualitatively. Only evaluating all of these entangled [3] factors jointly in context can yield insights into our participants' evaluations of both the differing models, as well as the assumptions of AQA.

These findings have a variety of limitations. Crucially, we cannot come to strong conclusions about the modeling work of any specific other authors, as our models were not implemented exactly like the papers which inspired them. We also cannot make claims about which of these models is best aligned with "human" preferences, as our participants were not a representative sample of a real-world population.

5.3 Recommendations for ML/UX Collaboration

To return to the scenario from the introduction: what should Alice and Brenda do to resolve their conflicting results? Ultimately,

we recommend approaching the ML problem statement as a user-centered design task, prioritizing the perspective of real human users. Early in the development process, before collecting training data, practitioners should engage in a user-centered situated evaluation process including prototype design, engineering and UX evaluation integrating both UX and ML practitioners. Even if the ML model used for prototyping is not similar in performance to a final model, the process of designing around a potentially inaccurate component can be an informative exercise: ML practitioners can learn how impactful different failure modes are for user experience and adjust their problem statement and evaluation metrics to compensate. Similarly, UX practitioners can adjust their mental models [4] of how an eventual ML system will work in context, allowing them to design realistically around the performance of an existing model, rather than their imagined ideal model behavior.

Additionally, our findings contribute three key insights to this conflict between evaluation standards:

(1) *Model users' expectations:* We found that participants' evaluations of the various ML models were mediated through their prior experience with photography, which shaped their expectations for our application's behavior. When developing user-facing ML-based tools, we recommend taking user perspectives and mental models into account when defining a problem statement. For example, rather than try to model photo aesthetics in general, data collection and modeling for our application could focus on the types of photos that a specific user population expects to take. In our case, that means gathering labeled data to model the aesthetic standards of bird photographers or party photographers, who have very specific expectations, rather than all users in general, and making those limits explicit.
(2) *Foreground Subjectivity:* In machine learning, there is a tendency to reduce the complexity of human judgment to an instance/label pair and reduce human inter-subjective variation to noise. When approaching fundamentally subjective problems, we encourage practitioners to foreground user experience and develop performance metrics that prioritize factors impacting user experience, rather than only measuring accuracy or a similar objective performance metric.
(3) *Design Legible Interfaces for ML Features:* When approaching subjective problems, users may be reluctant to believe that an AI system is actually computing the concept in question. Designers should not treat the ML components of these systems as black boxes or oracles. Instead, we encourage the development of legible models and interfaces, where the system behaves in a consistent manner that reveals the designers' interpretation and operationalization of the subjective concept. A legible interface does not need to be highly technical and transparent; it only needs to behave consistently and make the reasons for the system's behavior visible to users.

We have several recommendations for future situated evaluation studies. First, we found that a semi-structured approach allowed the experimenter's off-hand remarks and follow-up questions to influence participant behaviors. We recommend using a more structured scripted format in a consistent environment with consistent tasks to reduce this bias. Second, to eliminate the confounding factor of different model output distributions, we recommend ensuring that all models run at the same temporal frequency and are scaled to fill the output space as consistently as possible.

Finally, while we can center user perspectives in evaluation, we cannot escape the implicit way that computer science shapes problems, dataset and model development [13]. In other words, no matter how much feedback participants can offer, everyone will always be discussing problem statements, models and datasets developed by computer scientists. A more thoroughly feminist approach to subjective tasks like AQA would empower users to adjust the behavior of the ML components themselves to match their intuitions or reject these components entirely. We encourage future work on customizable, few-shot learning to increase the flexibility of ML solutions to subjective problems so that both designers and users can better model their own ways of seeing.

References

[1] [n. d.]. Meitu Photo Editor and AI Art on Google Play. https://play.google.com/store/apps/details?id=com.mt.mtxx.mtx retrieved May 15th 2023.
[2] [n. d.]. VSCO: Photo & Video Editor on Google Play. https://play.google.com/store/apps/details?id=com.vsco.cam retrieved May 15th 2023.
[3] Derya Akbaba, Lauren Klein, and Miriah Meyer. 2024. Entanglements for visualization: Changing research outcomes through feminist theory. *IEEE Transactions on Visualization and Computer Graphics* (2024).
[4] Robert B Allen. 1997. Mental models and user models. In *Handbook of human-computer interaction*. Elsevier, 49–63.
[5] Jeffrey Bardzell and Shaowen Bardzell. 2013. What is" critical" about critical design?. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3297–3306.
[6] Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a feminist HCI methodology: social science, feminism, and HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 675–684.
[7] Shaowen Bardzell, Jeffrey Bardzell, Jodi Forlizzi, John Zimmerman, and John Antanitis. 2012. Critical design and critical theory: the challenge of designing for provocation. In *Proceedings of the designing interactive systems conference*. 288–297.
[8] Kum-Kum Bhavnani. 1993. Tracing the contours: Feminist research and feminist objectivity. In *Women's Studies International Forum*, Vol. 16. Elsevier, 95–104.
[9] Charles Bigelow. 2019. Typeface features and legibility research. *Vision research* 165 (2019), 162–172.
[10] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. Pine Forge Press.
[11] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *ECCV*. Springer, 288–301.
[12] Nicholas Davis. 2013. Human-computer co-creativity: Blending human and computational creativity. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 9. 9–12.
[13] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society* 8, 2 (2021), 20539517211035955.
[14] Remi Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554* (2021).
[15] Remi Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399* (2020).
[16] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.
[17] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308.
[18] Anthony Dunne and Fiona Raby. 2013. *Speculative everything: design, fiction, and social dreaming*.
[19] Hamid Reza Ekbia. 2008. *Artificial dreams: the quest for non-biological intelligence*. Vol. 200. Cambridge University Press Cambridge.
[20] Michele Elam. 2022. Signs taken for wonders: AI, art & the matter of race. *Daedalus* 151, 2 (2022), 198–217.
[21] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of advanced nursing* 62, 1 (2008), 107–115.
[22] Jodi Forlizzi, John Zimmerman, and Erick Stolterman. 2009. From design research to theory: Evidence of a maturing field. *Proceedings of IASDR* 9 (2009), 2889–2898.

[23] Andrew Gambino, Jesse Fox, and Rabindra A Ratan. 2020. Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication* 1 (2020), 71–85.

[24] Barney G Glaser and Anselm L Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

[25] Samuel Goree. 2021. What does it take to cross the aesthetic gap? the development of image aesthetic quality assessment in computer vision. In *Proceedings of the 12th International Conference on Computational Creativity*.

[26] Samuel Goree, Weslie Khoo, and David J Crandall. 2023. Correct for Whom? Subjectivity and the Evaluation of Personalized Image Aesthetics Assessment Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[27] Elizabeth Gray. 2018. Understanding Auto ISO in Photography. https://photographylife.com/what-is-auto-iso, retrieved May 18th 2023.

[28] Leif Hancox-Li and I Elizabeth Kumar. 2021. Epistemic values in feature importance methods: Lessons from feminist epistemology. In *proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 817–826.

[29] Donna Haraway. 1988. The Science Question in Feminism. *Feminist Studies* 14, 3 (1988), 575–599.

[30] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. In *Feminist theory reader*. Routledge, 303–310.

[31] Mark Harman. [n. d.]. OpenCamera. https://opencamera.org.uk/.

[32] Wolf Hauser, Balthazar Neveu, Jean-Benoit Jourdain, Clément Viard, and Frédéric Guichard. 2018. Image quality benchmark of computational bokeh. *Electronic Imaging* 2018, 12 (2018), 340–1.

[33] 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.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[35] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding visual explanations. In *Proceedings of the European conference on computer vision (ECCV)*. 264–279.

[36] Aaron Hertzmann. 2022. The choices hidden in photography. *Journal of Vision* 22, 11 (2022), 10–10.

[37] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. " Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[38] Bjørn Karmann. 2023. Paragraphica – Context to image (AI) camera. Retrieved from https://www.creativeapplications.net/objects/paragraphica-context-to-image-ai-camera/ on 6-8-23.

[39] Pariya Kashfi, Agneta Nilsson, and Robert Feldt. 2017. Integrating user experience practices into software development processes: Implications of subjectivity and emergent nature of ux. *PeerJ Computer Science* 3, e130 (2017).

[40] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In *CVPR*, Vol. 1. IEEE, 419–426.

[41] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[42] Alexandra Kirsch. 2017. Explain to whom? Putting the user in the center of explainable AI. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2017)*.

[43] Lauren Klein and Catherine D'Ignazio. 2024. Data Feminism for AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 100–112.

[44] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*. Springer, 662–679.

[45] Angelie Kraft and Eloïse Soulier. 2024. Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1433–1445.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *NeurIPS* 25 (2012), 1097–1105.

[47] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[48] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.

[49] Congcong Li, Alexander C Loui, and Tsuhan Chen. 2010. Towards aesthetics: A photo quality assessment and photo selection system. In *Proceedings of the 18th ACM international conference on Multimedia*. 827–830.

[50] Cynthia CS Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. 2018. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. *Explainable and interpretable models in computer vision and machine learning* (2018), 197–253.

[51] Joseph Lindley, Haider Ali Akmal, Franziska Pilling, and Paul Coulton. 2020. Researching AI legibility through design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[52] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. 2014. Rapid: Rating pictorial aesthetics using deep learning. In *ACM Multimedia*. 457–466.

[53] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[54] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.

[55] Meena Devii Muralikumar and David W McDonald. 2024. Analyzing Collaborative Challenges and Needs of UX Practitioners when Designing with AI/ML. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–25.

[56] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*. IEEE, 2408–2415.

[57] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[58] William Odom, Mark Selby, Abigail Sellen, David Kirk, Richard Banks, and Tim Regan. 2012. Photobox: on the design of a slow technology. In *Proceedings of the designing interactive systems conference*. 665–668.

[59] Will Orr and Edward B Kang. 2024. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1875–1884.

[60] James Pierce and Eric Paulos. 2015. Making multiple uses of the obscura 1C digital camera: reflecting on the design, production, packaging and distribution of a counterfunctional device. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2103–2112.

[61] Cameron Pugach, Helmut Leder, and Daniel J Graham. 2017. How stable are human aesthetic preferences across the lifespan? *Frontiers in human neuroscience* 11 (2017), 289.

[62] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. 2017. Personalized image aesthetics. In *ICCV*. 638–647.

[63] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John A Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C North, and Daniel A Keim. 2017. What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing* 268 (2017), 164–175.

[64] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. 2018. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & quantity* 52 (2018), 1893–1907.

[65] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

[66] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[67] Ramya Srinivasan. 2024. To See or Not to See: Understanding the Tensions of Algorithmic Curation for Visual Arts. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 444–455.

[68] Susan Leigh Star. 1999. The ethnography of infrastructure. *American behavioral scientist* 43, 3 (1999), 377–391.

[69] Sherry Turkle, Cynthia Breazeal, Olivia Dasté, and Brian Scassellati. 2006. Encounters with kismet and cog: Children respond to relational artifacts. *Digital media: Transformations in human communication* 120 (2006).

[70] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[71] Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. 2013. Hoggles: Visualizing object detection features. In *Proceedings of the IEEE International Conference on Computer Vision*. 1–8.

[72] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420* (2023).