

“Impressively Scary:” Exploring User Perceptions and Reactions to Unraveling Machine Learning Models in Social Media Applications

Jack West
University of Wisconsin-Madison
Madison WI, USA
jwwest@wisc.edu

Bengisu Cagiltay
University of Wisconsin-Madison
Madison WI, USA
bengisu@cs.wisc.edu

Shirley Zhang
University of Wisconsin-Madison
Madison WI, USA
hzhang664@wisc.edu

Jingjie Li
University of Edinburgh
Edinburgh, UK
jingjie.li@ed.ac.uk

Kassem Fawaz
University of Wisconsin-Madison
Madison WI, USA
kfawaz@wisc.edu

Suman Banerjee
University of Wisconsin-Madison
Madison WI, USA
suman@cs.wisc.edu

ABSTRACT

Machine learning models deployed locally on social media applications are used for features, such as face filters which read faces in-real time, and they expose sensitive attributes to the apps. However, the deployment of machine learning models, e.g., when, where, and how they are used, in social media applications is opaque to users. We aim to address this inconsistency and investigate how social media user perceptions and behaviors change once exposed to these models. We conducted user studies (N=21) and found that participants were unaware to both *what* the models output and *when* the models were used in Instagram and TikTok, two major social media platforms. In response to being exposed to the models' functionality, we observed long term behavior changes in 8 participants. Our analysis uncovers the challenges and opportunities in providing transparency for machine learning models that interact with local user data.

CCS CONCEPTS

• Security and privacy → Social aspects of security and privacy; • Human-centered computing → Empirical studies in HCI.

ACM Reference Format:

Jack West, Bengisu Cagiltay, Shirley Zhang, Jingjie Li, Kassem Fawaz, and Suman Banerjee. 2025. “Impressively Scary:” Exploring User Perceptions and Reactions to Unraveling Machine Learning Models in Social Media Applications. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26-May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3706598.3713256>

1 INTRODUCTION

Social media apps have become integral to modern life, with 4.95 billion users connecting through mobile and web platforms [101]. Billions of people use mobile devices to access social media, with

Instagram and TikTok boasting over one billion [40] and five billion [97] downloads, respectively. One of the key factors to their popularity is content tailored to individual preferences. As innovators in the technology space [72, 104], social media companies are leaders in recommendation algorithms [49, 103] and machine learning (ML) models [68, 70]. To train these models, social media companies utilize user data [64].

Artificial Intelligence/ Machine Learning (AI/ML) models in social media *already* have a multi-billion dollar market size and are expected to continually grow [65, 73, 106]. AI/ML advances are happening at an unprecedented pace [96, 102], limiting users' access to accurate and up-to-date information and impacting companies to understand how and when AI is employed [51, 76]. In an article written in Meta's Transparency Center [69], they describe how their AI/ML models are used in their algorithm, “*The AI system considers a variety of **input signals** about each post. These signals might include who created the post or the type of content in the post. A lightweight model is run to select approximately 500 of the most relevant posts.*” Transparency statements from companies seldom explain *when* they apply these algorithms and *what* specific input signals they take into account.

While social media companies may not be legally obligated to be fully transparent about their algorithms (e.g., due to intellectual property concerns), the lack of transparency propels users to rely on their perceptions to explain what they see on social media [80]. Documenting and understanding these algorithmic perceptions is a well-established field in the HCI community [21, 23, 24, 37, 74, 82, 100, 122]. Prior works about social media algorithm perceptions focus on advertisements [86, 89], social media feeds [23, 24, 33], misinformation [15, 16, 116], and several other social media features relying on algorithm decision making [19, 21, 37]. Prior studies explore models based on patents [2] or focused on feed curation [31], and they have not explicitly exposed the functionality of authentic models to study participants. As a result, there is a gap in understanding the actual user reactions and perceptions towards *authentic* models employed by social media.

Here, we leverage recent advances in reverse engineering of mobile apps to enable user interactions with authentic AI/ML models from Instagram and TikTok [112]. With access to these authentic models, we aim to explore user perceptions and behavior changes

when exposed to transparent AI/ML functionality. In particular, we ask the following research questions:

- **RQ1:** What are social media users' *prior understandings and perceptions* toward the usage of ML/AI within social media applications?
- **RQ2:** What are users' *immediate reactions* to and understandings from our intervention, which reveals how ML/AI models use and analyze local data on Instagram and TikTok?
- **RQ3:** How do users' understanding of ML and associated *behaviors change* after learning about authentic models deployed by Instagram and TikTok?

To address these research questions, we conducted an exploratory user study with 21 social media users and evaluated how their perceptions and behaviors change once exposed to authentic AI/ML models. In our study, participants interacted with the models and explored *what* and *how* their data is being processed locally by social media companies.

First, we conducted an interview to understand the participants' *prior* perceptions and behaviors. We then demonstrated a basic example of computer vision using Google's Teachable Machine [14] and discussed concepts of computer vision. Second, we demonstrated two computer vision models (from Instagram and TikTok) and allowed the participant to interact directly with the model. A novel contribution of our work is the use of authentic AI/ML models and classifying participant data in real-time. To address the final research question, during the interview, we asked participants if they thought any behaviors may change after seeing the models. We then distributed a follow-up survey two weeks after the last participant to see if any reported behavior changes persisted.

Our findings reveal a mismatch between participants' expectations and how Instagram and TikTok actually employ AI/ML. They expected social media platforms to employ AI/ML on data they intentionally provide to the apps, such as posts and likes. However, direct exposure to authentic vision models heightened concerns around data collection practices, particularly regarding demographic information. These concerns depended on how participants interacted with the app. Participants reacted negatively to TikTok's model analyzing their camera frames without notification, explaining that it felt *non-consensual*. On the other hand, participants felt uncomfortable with the abundance of data Instagram extracted from their images (over 500+ different abstract concepts from a single image). They were confused about what Instagram's data would be useful for as the amount was *overwhelming* and the data's relevance to the app was *unclear*. While participants exhibited varying attitudes about the local processing of their data, they all expressed interest in increased transparency from social media apps. However, they had different preferences regarding what information the apps should make available, such as AI/ML inputs, outputs, purposes, timing, and controls. Further, our work explores user-reported *short-term* and *long-term* change. Interestingly, two weeks after the conclusion of our study, eight users (out of 17 who filled out our post-study survey) reported a change in their behaviors.

2 RELATED WORK

Our work aims to fill knowledge gaps in several areas in the HCI community. Specifically, we address gaps in AI/ML algorithms in social media, XAI and transparency, user perceptions towards algorithms in the context of social media, and user privacy perceptions.

2.1 Understanding of AI/ML Algorithms

A large body of work studies users' understandings and reactions to AI/ML [2, 5, 41, 52, 61, 78, 88, 118, 120, 121]. The existing works suggest that participants have mixed reactions, which depend on context. For example, Ashktorab et al. [5] performed a large-scale user study where participants were paired with AI in a game where the parties could each have the role of question giver or answering the questions. Participants' likeability towards AI significantly dropped when the participant needed to depend on the AI to guess questions correctly. However, participants reacted more positively when the AI was the question giver. Ashktorab showed that *context matters* when understanding reactions to AI. To further the idea that context impacts user reactions to AI, Kapania et al. [52] interviewed citizens of India and found that people were overwhelmingly positive towards AI decisions, as they felt they would be more correct than human decisions. However, users tend to have negative attitudes toward the usage of AI in algorithms [2, 25, 81, 121]. Zhang et al. [121] interviewed several users about their attitudes toward computer vision monitoring. They found that users were most likely to be somewhat uncomfortable (36%) with video analysis done on them. We expand on these works by establishing what specific details about modern AI/ML users do not like about model deployment. We show that AI/ML sentiments also apply to *implementation method*, not just results as indicated by these works.

2.1.1 Understanding AI/ML Algorithms in Social Media. Exposing participants to underlying algorithms in social media has been extensively studied [2, 32, 33, 36, 50]. Feng et al. [36] found that people *did not trust* the algorithm to give them the best content possible. They thought that an algorithm would misunderstand or miss simple nuances. Jahanbakhsh et al. [50] found participants' agreement rate of AI curated content fell once they were told the posts were selected by an AI. Feng et al. and Jahanbakhsh et al. show that participants lack trust in AI personalization. Andalibi et al. [2] demonstrated users' observed reactions to emotion recognition ML models in the context of social media. They found that several users had overwhelmingly negative reactions toward having their emotions analyzed, as they felt it was too personal. Our work builds upon these and differentiates itself in three ways: first, we present participants with real-world models of how AI/ML models are used for image classification, offering a more practical demonstration. Second, our approach is specifically tailored to social media platforms such as Instagram and TikTok, grounding our discussion in users' direct experiences. Third, we present AI/ML models to users separate from any algorithms, focusing on how participants felt about it being used in a more general sense.

2.1.2 Ecological Validity in AI/ML Research In Social Media. While there exists work that utilizes reverse engineering techniques for security studies on smartphones [7, 107] similar methods have yet to be applied to AI/ML in social media. Research in social media

has inherent trade-offs on ecological validity [8, 87] due to the opaqueness of the algorithms they explore. Because of this limited information, similar studies relied on patents [2, 85] or simulations [15, 86] to reconstruct AI/ML functionality. As a result, prior works that studied image analysis of users [121] lacked the prior experience and private reflections that our study portrays. Works that specifically deal with authentic algorithms [33] also exhibited the same reflective thought process. Participants reacted more strongly to models operating *within the camera* without their knowledge. Authentic models within social media apps offer two unique nuances: personal (1) reflections and (2) reactions.

2.2 AI Explainability and Transparency

AI explainability [27, 29, 36, 55, 63, 83, 111, 117, 119] (XAI) is a body of work that explores how AI methods should be explained to user of an application. AI transparency [4, 30, 34, 35, 60, 63, 90–92] discusses the impact of applying and methods for implementing transparency for AI. Our work highlights aspects of both topics. Liao et al. [63] described how AI systems should inform users about how and why their personal information is collected. They found that more transparency could translate into a higher sense of user trust. Further, AI transparency and explainability are only becoming more relevant with the HCI community as the European AI transparency laws [77] are implemented. El et al. [30] hosted a workshop discussing potential directions and important questions for the HCI community regarding both transparency and explainability. While the work's context surrounds generative AI, we found that some of the broader questions can also be applied to other types of AI/ML models. Ehsan et al. [28] demonstrated that explainability is relative to the person using the tool. They claim that those who are more comfortable with AI are able to interpret the tools better. Our work contributes to the ongoing discussion around AI explainability and transparency by addressing user understanding in the context of social media. While prior studies highlight the importance of transparency in increasing user trust [63], explainability varies based on user's familiarity [28], we extend the research area by exploring how the social media environment affects AI explanations.

2.3 AI/ML Privacy

Recent studies have explored new privacy and trust concerns around quickly advancing AI/ML capabilities [6, 17, 56, 59, 108]. Vereschak et al. [108] found that for AI/ML models to be trusted, the models must have positive expectations from the user. Vereschak et al. also argued that this trust depends on the provider of the model. Asthana et al. [6] explored the question of how participants felt about certain model inferences on their data. They found that participants were least comfortable with data being collected on *protected* attributes such as race. Our work examines the privacy perceptions towards local machine learning models and defines new problems that stem from local AI/ML processing.

3 METHODS

We conducted and analyzed semi-structured interviews with adult social media users (N=21) in the U.S. We include our recruitment

material, screening form, interview protocol, and post-interview survey in an OSF repository for open access.¹

3.1 Participants and Recruitment

We recruited participants through a screening survey distributed using our institution's mass email service. We conducted interviews with 21 participants (see Table 1) over a one-month time period from June 21st to July 25th, 2024. All interviews were performed in-person at a private location on our institution's campus, and consent was obtained from participants at the beginning of the interview. Upon completion of the study, participants received \$30. The study was approved by our Institution's IRB. The length of each interview was between 48 to 82 minutes, averaging 64 minutes.

Over 849 people filled out the screening survey, with 831 being qualified for the interview. Our screening survey asked respondents if they were comfortable with an in-person interview, their age, and to list what social media they currently use. The exclusion criteria included respondents who were not comfortable with an in-person interview, below the age of 18, or did not use Instagram or TikTok. Individuals who could not attend in person were excluded due to parts of the study requiring interaction with the machine learning models. However, P18 was selected even though they did not actively use Instagram or TikTok. We decided to include them in this study as they used Facebook and, therefore, indicated to us that they were familiar with the platform. They also informed us, during the interview, that they wanted to get an Instagram to monitor their child's performance in sports.

In the survey, we collected information including how many hours a week participants used social media, whether they use social on their cellphones, their social media usage patterns (hours per week, platform of choice, commonly used features, etc.), and demographics. We then selected participants who used their phones for social media, balanced age, self-identified gender, tech-savviness ([98]), and their frequency of social media use.

To gauge tech savviness, we presented two statements: (1) "People consider me to be tech savvy." and (2) "I consider myself tech savvy." Individuals rated these questions on a 5-point Likert scale.

3.2 Semi-Structured Interview Procedure

As shown in Figure 1, our semi-structured interview included five phases that focus on the following topics.

Establishing Initial Understandings and Assumptions Towards AI/ML and Social Media. We explored each participant's initial sentiments and perceptions regarding social media. Specifically, we inquired about their favorite apps and why they use them. We also asked about the content they consumed and their usage habits (posting, how they used social media, etc.). Following the social media questions, we ask them about their understanding of ML. These questions establish (1) if they had heard of ML before and (2) to explain what they think ML means. If the participants had never heard of ML, we asked them if they had heard of AI. They were then tasked with defining the difference if the participant had heard both ML and AI prior to the interview.

¹https://osf.io/m4nru/?view_only=29ec93cc559d46af94d22ea9b100a4df

Table 1: Participant demographics and tech savviness (Self reported, rated on a 5-point scale)

Participant ID	Identified Gender	Age Bin	Highest Education Completed	Tech Savviness
P1	Male	26-35	PhD	4
P2	Female	26-35	Masters Degree	3
P3	Female	26-35	Masters Degree	3
P4	Male	18-25	Undergraduate Degree	3
P5	Female	36-45	Masters Degree	4
P6	Male	36-45	Masters Degree	4
P7	Male	36-45	Masters Degree	3
P8	Non-Binary	18-25	High School	3
P9	Female	18-25	Masters Degree	2
P10	Non-Binary	26-35	Masters Degree	5
P11	Male	46+	Undergraduate Degree	5
P12	Male	26-35	Masters Degree	4
P13	Female	46+	Undergraduate Degree	4
P14	Male	18-25	Undergraduate Degree	2
P15	Female	18-25	Masters Degree	4
P16	Male	46+	Undergraduate Degree	4
P17	Non-Binary	26-35	Undergraduate Degree	2
P18	Female	46+	Undergraduate Degree	2
P19	Male	46+	Masters Degree	5
P20	Male	18-25	High School	3
P21	Male	26-35	Masters Degree	2

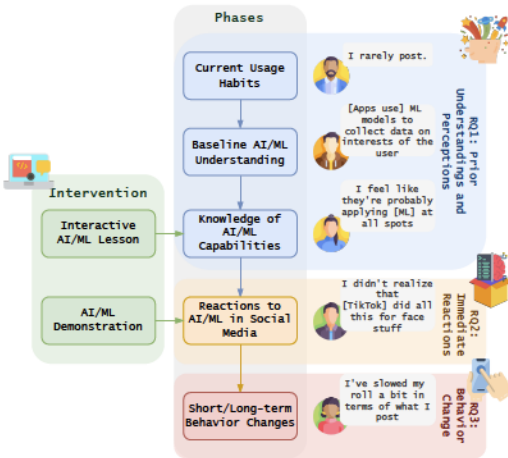


Figure 1: A high-level overview of our semi-structured interview, including example responses. The interview consists of five phases: (Phase 1) We explored participants’ past interactions with their social media apps. (Phase 2) We identified participants’ current understandings and assumptions towards AI/ML. (Phase 3) We conducted a learning activity, including a high-level example of computer vision, and discussed its capabilities. (Phase 4) We demonstrated the AI/ML models for TikTok and Instagram and observed their immediate reactions. (Phase 5) We surveyed short-term and long-term changes in participant behaviors.

Defining Computer Vision and Clarifying Misconceptions About AI/ML. We transitioned to a learning activity to support users’ understanding of general capabilities of AI/ML. Using Google’s Teachable Machine [14] we guided participants through a simple example of training and using a computer vision model. To avoid common pitfalls from misconceptions created by machine teaching tools demonstrated by Hong et al. [43] we taught the participant

ourselves avoiding an entirely automated process. We designed a brief lesson where the participant trains a model that detects when they are in the frame and when they are not in frame. To facilitate the process we provided props to the participant to test different classifications. Once the new model was trained, we asked questions regarding their understanding of AI/ML based on the new content they practiced. This phase allowed for each participant to engage in more in-depth conversations relating to AI/ML concepts.

Asking Participants “When” They Think AI/ML Happens in Social Media . We defined AI/ML expectations for social media applications in two ways: *when* AI/ML should be used on their images and *what* data should be observed from their images. We defined four instances of when the model could be used: “Opening Photo”, “Editing Photo”, “Face Filters”, and “After Posting”. Each scenario was demonstrated to the participant using Instagram on our study’s Android phone. For each instance, we asked participants when they believed AI/ML was happening and why.

AI/ML Model Demonstration and Participant Interaction. We demonstrated the authentic AI/ML models in real time. These models were adapted from the methods from West et al. [112] (for details see Sec 3.3). We first showcased TikTok to the participant and followed by Instagram. The interviewer first showed how the AI/ML model works with their own face using the app. We then asked the participant to observe the output (example output in Figure 2) and describe whether the model lived up to their expectations, if they were comfortable with the what they observed, and if they would’ve liked to know about this before using the application. The participant then explored the model output themselves. While exploring the models, participants were informed that we were unaware of the purpose for the model outputs.

Assessing Short-Term and Long-Term Behavior Changes. We asked participants general questions comparing their experiences with the two applications. During this process we asked about their feelings, reactions, and thoughts about what they have been shown. We also asked them to decide which method of AI/ML they preferred



Figure 2: Example output that participants would observe during the interview. Instagram displays over 500 different decimals and concepts which are ordered from largest decimal to smallest. If the decimal value is closer to one than it Instagram perceives that concept in the image; if the value is closer to zero the concept is not in the image. TikTok observes age and gender age is an exact estimate and gender seems to be a probability. Gender is a probability called *boy_prob* when the decimal is closer to one it assumes a male gender identity if its closer to zero it assumes a female gender identity. Note that here we use a image generated by GPT-4o mini [75] for illustrative purposes instead of the input picture from real people in respect of individual privacy.

and why. Two-weeks after the last interview, we sent out a brief post-study reflections survey. The survey inquired about if any behavior changes since the interview and/or if sentiment changes towards social media. We sent the survey to all 21 participants and 17 responded.

3.3 Extracting AI/ML Models from Instagram and TikTok

With the increase in technological abilities of mobile devices, there has been an industry-wide push to host local ML models [53]. Hosting ML models locally offers several benefits, such as increased user privacy, low latency, and algorithm personalization. ML models of any modality are loaded using either a third-party library like Pytorch [79] or a custom ML library specific to the app. Once loaded, ML models operate as they would on a server, existing as large blocks of memory performing inferences on input data points. However, since the beginning of this migration, security researchers developed methods to extract, steal, and corrupt local ML models [13, 22, 45, 46, 84, 99, 112]. To access ML models, three conditions must be met: (1) the researchers must have access to a rooted Android device, (2) the app containing the ML model must be statically analyzed, and (3) the app must be dynamically analyzed. To fulfill the first condition, we acquired a rootable Pixel 4 device and followed a guide to root the device [42]. To accomplish the static and dynamic analysis requirements, we used the same methodology described in West et al. [112]. We first decompiled the apps using JADX [93] and Ghidra [1] to search for keywords that relate to machine learning. Once relevant functions were identified, we then dynamically analyzed the apps from the functions of interest using Frida [38]. However, for TikTok, the described method does not work due to its reliance on dynamically defined functions.

West et al. provide the function name (*MessageCenter.postMessage*) where the AI/ML output is displayed, which we use in this work.

3.3.1 *Our Understanding and Assumptions of The Models.* Given that our method for accessing TikTok and Instagram’s models was through reverse engineering, we have limited insight into how the models function, why, and what the information is collected for. We found that TikTok’s AI/ML model is built into the camera driver, meaning that all camera frames are directly handed to the model while the camera is open. The model is *always* active, a filter does not need to be enabled, and we did not enable any filters when demonstrating the models to participants. The data that we show participants are the result of the outputs being passed *out* of the camera driver and into the Java layer through a *callback* function. The data is then written to an encrypted file. Instagram’s model executes when the user selects an image to upload as a Reel. This selected image *does not need to be uploaded*. We found that the model only executes upon the selection of a photo. We *cannot* assume the data’s purpose due to technical limitations (see Section 5.3). We also checked Instagram and TikTok’s privacy policies [48, 105] and could not confidently connect their privacy policy terms to either model we study in this work. Thus, we decided *not to inform* participants of our reverse engineering endeavors as it could lead to a bias. During our interviews, we informed participants that we were unaware of where the data went or what it was used for.

3.4 Data Analysis

To analyze the semi-structured interview using reflexive thematic analysis [11, 66, 114] (RTA). RTA acknowledges the subjective and interpretive roles of researchers, viewing them as active contributors to meaning-making rather than objective coders. This approach was particularly beneficial given our team’s diverse expertise in cybersecurity, machine learning, and HCI; it allowed each researcher’s unique perspective to inform and shape the development of codes and themes. As a result, inter-rater reliability measures were not applicable within this reflexive framework [12]. To transcribe the audio, we use the tool noScribe [26], which transcribes audio locally without third-party APIs. All transcriptions are then verified by human reviewers. Three researchers coded the first three interviews separately and had productive and iterative discussions pertaining to developed themes. The researchers held periodic meetings to discuss codes, themes, and findings. The codes and themes were identified and developed in this process, and the disagreements were resolved collaboratively. The codebook was finalized by the sixth meeting, after which we continued discussing and analyzing our results. All other transcripts were then coded using that codebook by a single researcher. We also collected binary and open-ended answers about how participants’ behavior changed in the post-interview survey. In the analysis, we clustered the participants into three categories based on their answers: changed behavior, did not change behavior, and planned to change behavior. The final themes and survey results are reported in Section 4.

4 RESULTS

We identified several recurring themes from our qualitative analysis, which we present in the following seven categories: (1) participant

assumptions towards AI/ML and social media, (2) participants' usage behaviors, (3) perceptions around model deployment, (4) AI/ML privacy, transparency, and trust within social media, (5) comparisons between the two models, (6) data collection assumptions, and (7) reactions to the AI/ML models within social media.

4.1 Participant Assumptions and Initial Perceptions on AI/ML and Social Media

Among initial perceptions and understandings toward AI/ML and social media, we noticed three recurring themes highlighting participants' assumptions: (1) app interactions (likes, follows, searches, etc.) were assumed to be the main inputs to AI/ML for advertising and recommendation algorithms, (2) vision models were assumed to be used as a source to infer demographics in the context of social media, and (3) participants were unaware of the speed of modern computer vision models.

4.1.1 The Purpose of AI/ML in Social Media is for Advertising and Recommendations. Our participants shared similar assumptions on how social media used information, mainly focusing on their content interests. They all believed that social media learned about them through their interactions within the app. P4 explains

"I think kind of the obvious [ML method] is they're using machine learning models to collect data on interests of the user. So understanding or collecting data on what they interact with, whether or not they like something, comment on something, share something, follow more pages that are related to a specific bucket of content."

P20 expressed similar assumptions, stating that apps will probably look for words, topics, and accounts people interact with. Participants assumed that data being observed by social media had the purpose of improving the experience (N=6), for profit (N=13), or just to collect data (N=3). P3 explained their skepticism of free social media apps earn profits by collecting data that *"is giving them profits in some way, shape or form."* Several participants assumed that *user interactions* were collected to provide advertisements and/or recommendations all to increase profit (N=8). We demonstrate that participants assumed *user interactions* were the inputs to their advertising and recommendation algorithms.

4.1.2 Assumed Purpose of Computer Vision Models in Social Media. We asked participants, during the lesson, what they thought computer vision models could be used for within social media, and two main themes developed: (1) to offer features or improve algorithms (N=4) and (2) to infer more information about the user (N=15). When asked how they thought a social media company could use a computer vision model, P9 assumed social media would want to find demographic data because they felt Instagram already *"have [P9] pretty pinned down."* Participants assumed that, for a successful personalization algorithm, social media apps would need to know more about the user. P15, when asked about the purpose of computer vision models in social media, stated,

Probably gender, age, location, geographic location. I think part of their algorithms or models would especially now be driven toward what would cause me to purchase things or what they should advertise to me.

P15's statement subtly implies that computer vision models may be able to identify several demographic indicators and that data would best be used for advertising. Those who assumed that the models existed to improve their experience thought that the models did so by collecting demographic information or abstracting engagement visually. P16 assumed that the data collection was necessary for the app as vision models could help everyone they explained, *"I mean, if [the models] helps man detect and helping people, I mean, that's beneficial."* P14 thought that Instagram's model could be used to detect explicit content, stating,

"So yeah, like nudity is one of them up there. It's probably like, you know, you kind of have to like regulate that."

P5 thought that vision models would be used to *"understand how people are interacting and to keep them interacting if they have high interactions with one visual topic, for lack of a better word."* Both groups acknowledged and were able to differentiate between app interaction models and vision data. Notably, because participants associated vision models with inferring demographics, most participants assumed that demographic estimation was why social media needed vision models.

4.1.3 The Model is Faster Than Participants Initially Assumed Possible. Participants who assumed that their images were being analyzed by the applications were still surprised by how *fast* the model collected data. P10 described how fast TikTok was when they were just *"on screen for like a second,"* and *"it was still able to pull up all this data."* P4 similarly expressed their surprise, stating

"It is still a little bit intimidating and whatnot, but I wouldn't put it past these things to be able to do it this fast. I've just never really thought about it too much. But yeah, I mean, I guess it is still surprising. It's impressive."

Overall, the model execution speed to some participants also felt invasive as they could not identify when it was happening. This highlights that participants assumed that AI/ML could be detectable if it was being used.

4.2 Participants' Usage Habits Influenced Their Reactions to the Models

We identified that participants' past usage habits on social media may create a bias that could shape their reactions to the AI/ML models. Participants introduce three primary reasons for using social media: leisure, social connections, and professional purposes. All participants (N=21) use social media for *leisure*, with some (N=3) indicating that leisure is their only reason for use. For instance, P6 said *"For general use, day to day, just for enjoyment. I doom scroll on TikTok quite a bit."* Most of the participants (N=18) used social media for a combination of leisure, social connections, and/or professional purposes. The second most common reason was to maintain *social connections* (N=16): P12 uses social media to keep up with friends and family, explaining,

"I come from [foreign country]...I wouldn't say it's a huge following. It's just that I've made friends there in this community there."

Lastly, a few (N=3) participants use social media for a *professional purpose*, such as P5, who has a social media to advertise their business, *"primarily about [their] art practice,"* despite their rare posting. Those who use social media as a source of leisure interacted with the camera *less* than those who utilized the apps for social connections and professionalism. The difference in participant *interactions with the camera will influence their short and long-term reactions to the models.*

4.3 Participant Perceptions Towards AI/ML Deployment

We explored participants' perceptions of when AI/ML should be utilized by social media applications. We noticed that participants disagreed on *when* AI/ML may take place.

4.3.1 When AI/ML Takes Place Is Not Clear to Participants. In the third phase of our interview, detailed in Section 3.2, we presented four options where AI/ML might take place for participants to choose from: (1) "Opening Photo", (2) "Editing Photo", (3) "Face Filters", and (4) "After Posting". We noticed *disagreement* amongst participants for "when" they assumed AI/ML occurred on social media apps. Half of our participants considered *AI/ML happens in all presented scenarios* (N=10), and the remaining participants (N=11) chose *a subset of scenarios they thought were likely*. Among those participants in the second group, only one participant (P16) selected "Opening Photo". Some participants in this group were weary of data being collected during the "opening photo" stage, P7 proclaimed *"Because opening the photo is in my device, it's still not posted and they won't receive the data from me."* Participants who selected all four options felt that the "Opening Photo" stage was unlikely. P3 states that they *"don't necessarily know about 'opening the camera.'" and feel like the apps are "probably applying at all spots."* P5 argues that this confusion correlated with a lack of transparency:

"I know [AI/ML is] there. Google sends me like 'oh six years ago', right? Like 'oh and this themed slide shows of'...which are like cat pictures where they're all together...they're clearly scanning and watching what is in my photo album and so, it's like, you know it's there."

Overall, we identified that *participants were confused on when AI/ML actually happened* which, some participants argued, that the lack of transparency caused the confusion.

4.4 Privacy, Transparency, and Trust in Social Media Applications

We found that lacking trust played a role in several assumptions toward data usage and local data processing. This lack of trust stems from social media's inferred business model: user data. Participants assumed that the outputs from the models are likely used for some underlying algorithm to further profits.

4.4.1 Participant Trust Influenced Assumptions Towards Local Processing. When we informed participants that we did not know what happens to the model outputs, we observed that several individuals (N=11) expressed two different reactions: (1) locally processing the data *would not* change their attitudes towards the models (N=6) and (2) local data processing *would change* their attitudes (N=5).

P20 expressed that they want to know because *"it's still using the technology that I own, because it's basing it off of the phone. It's using it in a way that I did not consent to, nor unless I'm lucky, I don't know what's happening."* But when local processing was further explained to P20, they stated, *"[local processing] makes it a little less malicious than I was perhaps making it out to be."* P20's malice was a result of their perceptions of the data's purpose, stating they assumed *"[social media apps would] likely just logging demographics and then maybe, you know, looking into like what they could boost based off of those demographics to."* Even when understanding that the readings would never leave the phone, they assumed that it would still be used for advertising locally. These feelings stem from a lack of *trust* in social media apps. When asked if P4's opinion of Instagram would change if the data did not leave the phone, they stated,

"So unless I know, like, I have definitive proof of knowing where that goes, and I just assume the worst. I'm a skeptic in that way. And so, you know, Instagram and TikTok, they could tell me that that's on my phone and on my local server or whatever, but I tend to not believe that."

P4 expresses a lack of trust towards the applications due to the monetary value of their data, stating, *"So part of me says it would be good for them just to delete that data and not have it be out there. But I know that that's not their business model."* P3 thought the data *must* be collected because *"Instagram is a free, free thing. So where are they getting the money from? So obviously, right, they're collecting data, and that data is giving them profits in some way, shape or form."* P3 and P4 both assumed that their data had monetary value and thus assumed that it would be used or collected. We identified that participants' lack of trust in social media influenced their opinions towards the apps processing local data. Interestingly, participants' lack of trust is *not* directed at the AI/ML models themselves but rather the applications deploying them.

On the other hand, P7 stated that if they knew their information did not leave the device, they would be fine with AI/ML analyzing their data. P7 *"assumed that social media is a business"* and our data was how they made profits, but *they must follow the law*. When asked if P7 would implement their hypothetical algorithm designed for social media, they stated,

"Ethically I wouldn't but... And legally I would... If it's not legally allowed to do, I wouldn't."

P14 expressed a similar sentiment by assuming that the models are likely disclosed in the *"terms of service"* without reading it. While P7 and P14 put their trust in the law, other participants felt agnostic about local data processing because they didn't know where it went. P12 said that their data had *"so much potential"* and thus did not full dismiss the possibility that their data was being collected. These results imply that only locally processed data may not raise trust in social media applications due to the assumed business model. Those who did trust local processing only did so when the trust was built on the law.

4.4.2 Participants Disagreed About How Much Transparency Should be Given to Users. All participants wanted more transparency about

AI/ML models used on their devices. However, there was disagreement about how much data apps should be transparent about. Participants exhibited one of three themes: (1) apps should detail as much information as possible (N=12), (2) apps should only provide limited information (N=5), and (3) the participants felt they already knew about data collection; thus transparency felt unnecessary (N=3). Most participants expressed curiosity about the models' purpose after being exposed to them. One example, P13, stated that "Not knowing what the purpose is for makes it harder to be accepting that it's happening." Others expressed a want to control the data. P5 explained that they, "would like control and not be able to give it information that I don't want to give it," when thinking about how their data could be used for advertising. However, P5 later acknowledges that it may render their algorithm useless, stating,

"But then what good is the algorithm? And what are we using this algorithm for?"

P5 expresses mixed feelings as they recognize that control could halt social media from learning about them but, at the same time, would hinder their algorithmic experience. Some participants wanted to know, but they did not want all of the information we provided. P11, for example, thought there was too much detail, especially regarding Instagram's model. P9 and P13 worried about the potential societal impact when revealing the outputs to the models. P13 assumed people would get side-tracked by the model, which would create a distraction. P9 assumed that people would attempt to optimize their scores, and it would drastically harm the next generation, stating,

"I mean, like people would be like, why is it my face 0.9 compared to Becky's face to one? Now I don't have as many followers because the score isn't higher. And how can I get it higher? And yeah. I think it could be detrimental for sure."

P9 reflected that users may define their self-worth by the outputs from the model. These assumed risks of increasing transparency are a direct result of the context of social media. The final notable opinion is that the participants knew this was happening and, therefore, felt it did not apply to them. For example, P16 said that they "kinda of figured" something like this was already happening and didn't see a need to be told because, to them, the fact the model existed was obvious. Overall, we identified that participants expressed multiple levels of comfort towards transparency.

4.4.3 Some Participants Did Not Expect the Camera to Analyze Them. Participants were confused about "when" a model was executed. P8 expressed surprise when the model was active, "I didn't realize that [TikTok] did all this kind of stuff for face stuff. I don't know. I just thought the camera just looked at you." P1 similarly reflected on how uncomfortable it felt knowing they had opened TikTok's camera in the shower before,

"That's quite surprising. Sometimes when I take a shower, I sometimes play [with my] cell phone. I mean I use my cell phone when I take a shower. Well, now that's... I'm naked"

These examples demonstrate the vulnerability and risk associated with AI/ML models use of passive camera stream.

4.5 Participants Compared Instagram and TikTok's Model Implementations

AI/ML models deployed by Instagram and TikTok analyze user images (Instagram) and camera frames (TikTok) on user mobile phones. Participants exhibited various reactions after experiencing how the TikTok and Instagram models operate: their reactions centered around the models' *inputs*, which is how images are provided to the model, and *outputs*, referring to the models' analysis of the given images. Participants shared that TikTok's model felt non-consensual because of how the images were delivered to the model. However, participants were uncomfortable with the amount of data that Instagram observed. We found that ten people preferred TikTok's model, while only seven people preferred Instagram's.

4.5.1 Comparisons Toward Model Input Methodology.

Participants Felt That TikTok's Model Was Creepy. Due to the fact that TikTok's model was active while the camera was open and undetectable to participants, they felt it lacked consent to analyze their images. When asked how they felt about local image analysis on their data, P20 angrily emphasized their disappointment of having "no way of knowing what it's picking up on, what it's storing...and what it's doing without [their] consent?!" P17 reflects on TikTok's model output, considering that "plenty of people might accidentally open their TikTok camera" without "thinking about giving data in that moment." Participants felt that because TikTok did not notify users about their data being collected, they felt TikTok was creepy, especially to P13, who stated, "So TikTok, your camera is open. You haven't even like recorded or posted and it's doing that? Oh, yeah. That's that's creepy!" P19 felt that TikTok was "Big Brother-ish"

"feels more like a place that nobody else deserves to be. Just me. Almost like somebody hacking into and watching you through your camera at home and you don't know it."

Overall, participants expressed negative perceptions upon learning how TikTok's model functions due to the lack of consent for data collection, and the limited transparency regarding their practices further exacerbated the participants' feeling 'creepy' and invasive.

Instagram's Model Felt Less Creepy. All participants had a notably better perception of Instagram's model implementation, and they considered it less creepy than TikTok's. P20 thought "Instagram model is more consensual because you have to take an image or look at an image that's already on your phone to activate the model." This suggests that with Instagram's model, there was a sense of control, which was missing from TikTok's model. P18 explained,

"Instagram is being a little more discerning and letting you do some selection before it runs through that filter and waits to see what you take a picture of before it turns this on."

These feelings were strong enough for several participants (N=7) to prefer Instagram over TikTok because of the extra perceived control. P18 thinks "TikTok seems more aggressive, but it doesn't seem like as radical a difference [between Instagram and TikTok]." When asked to clarify what they meant by 'aggressive,' P18 refers to TikTok's attempt to assess "who you are and what you might buy the second you turn on the camera." This extra control on Instagram

soothed some participants as they now knew *when* the AI/ML could happen, and they felt in control.

4.5.2 Model Outputs are Confusing and Misaligned with Expectations. The negative reactions to Instagram were shaped around the magnitude of data presented as part of the model's output. Participants felt the data was *random* as the sheer amount of data was difficult to comprehend. Some participants (N=8) tried to assign *assumed meaning* to the data by connecting their folk theories to the new data. Participants were *confused* as to why Instagram selected the values to assign to their images.

Instagram's Data was Too Confusing. The majority of participants were confused by the outputs. P19 described the labels as a "drunken mistake" and found "it very fascinating in several different ways." The sheer amount of data lead most, who did not assign meaning to the labels (N=8), to believe it was useless. P15 explains why they do not feel uncomfortable necessarily as "this doesn't mean anything." Whereas some participants (N=2) associated the data with negativity P13 summarizes how others might feel, by stating

"I'm still kind of neutral to negative about it because there's so many words that they are looking that don't apply at all. And so then again, I'm like, what's happening with this information? And then like even saying coming up with percentages that comes out of this many people do this. It's like, well, I'm not at a beach. And I don't know things like that."

P15 argues that the confusion is a product of not grasping where the data fits in an algorithm, and they "don't really have an explanation for why", when comparing the algorithms of TikTok and Instagram. TikTok's model appears to P15 that it infers attributes like gender and age which "are two driving factors of how people interact with media in general, not just social media;" while Instagram's model feels like "they're trying to understand more about what's in the picture." These reactions are likely rooted in the participants' initial understanding of the data's likely purpose.

Instagram's Data Felt Random. The most common reaction was that Instagram's data felt 'random' (N=19), as described by P3 who thinks "it is surprising in the randomness of these categories." Participants also noted that the data was confusing, as put by P12,

"I guess my question is why... I mean, there's more than 500 things in the world, right? So why these 500 categories?"

To the participants, this information was difficult to understand as P15 feels "like they're trying to understand more about what's in the picture, but [they] don't really have an explanation for 'why'." This reaction highlights a potential flaw with being transparent. When revealing the model outputs, the participants felt the data was *random*. The data feeling random lead to several assumptions about the data's purpose or assuming that the data had no purpose.

4.6 Participants Were More Likely to Accept Data Collection Practices When They Understood the Data

Participants were often able to make sense of TikTok's data. Some did not like the way the model was used without explicit consent

but felt that they *understood* why the data was being analyzed (N=8). P20 immediately assumed that TikTok's data collection, including demographics, was for advertisers or to "farm engagement" More participants preferred TikTok (N=10) over Instagram. This is likely due to TikTok collecting less data and participants feeling like they understood what the data was used for. When asked which model they were less comfortable with, P8 refers to Instagram as "it seems like there's all these different things that they're trying to categorize, which is a bit weird." P4 similarly contributes to this line of thought after comparing both applications

"I would be more likely to use TikTok because I think that it's collecting far less data and maybe, maybe less harmless data. That might be not a true statement. But yeah, just based on what you've seen. Yeah, I'll say because of less, less data."

Overall, the lack of transparency towards the model likely left people confused and speculating about the purpose of the data. Which, in turn, leads to further algorithmic theories or confusion about what this data may be used for.

4.7 AI/ML Model Reactions and How They Influenced Behaviors

We interviewed participants to see whether they thought their behavior would change in response to the intervention during and 2 weeks after the interview. 17 out of 21 participants replied to the post-study survey we sent to inquire about long-term behavior changes. We refer to the participants' answers immediately post-intervention as *short-term* changes, and we refer to the answers to the post-interview survey as *long-term* changes. We found that 6 participants expressed short-term change, and 8 expressed long-term change. Notably, 4 of these 8 participants had stated that they did not plan to make any long-term change.

4.7.1 Participants Inferred Meaning Toward Instagram's Data. Participants expressed one of two reactions with regard to Instagram's output: (1) they could infer use cases for the data (N=8), or (2) the data did not make sense (N=11). Those who gave meaning to the labels had mixed reactions based on their assumed purpose or consequences. For example, P10 was uncomfortable with the concepts related to location, "even if [they] don't post my face specifically." Similarly, P7 does not "want any company to get location from [them] just with a picture." These negative reactions focused on labels that they *personally* did not like. However, there were also positive feelings toward the potential uses of the data. For example,

"Like even though some of [the concepts] are goofy to me. They seem less. They seem more objective in a way that is less dangerous to get wrong."

P18 assumed that because the concepts were general if the AI/ML model were to get them wrong, there would be fewer consequences compared to TikTok. Due to the lack of transparency, participants were confused as to what the purpose of the data was.

4.7.2 Models Validated Prior Assumptions. After seeing the models, P8 felt vindicated in their decision to delete TikTok because they felt like "it was way too personal." Participants often referred to the models as "eye-opening," stating that some prior assumptions before

the interview were validated when observing the models. P13 felt like their thoughts have been confirmed after exposure stating *“this is like in your face. This is what’s happening.”* P9 shared a similar sentiment and already assumed that *“everything [they] hear about machine learning and data collection”* was happening. After an *“eye-opening”* experience of seeing the model, P11 discussed further assumptions if it operates at *“camera level”* or *“transmit level”* and combines data *“in a million different ways.”* These responses indicate that our intervention offered some context to participants’ prior perceptions about AI/ML use on social media. P5 speculated how Instagram’s model is used to keep people on the app,

“[The model is] useful for [Instagram] as far as their goals of keeping people on the app and how they see their end for capturing the data that would be useful for them as far as how people interact or how people use this information and being able to mine data of who’s interested in what and then being able to sell that data or, I don’t know, I don’t know. Big Brother’s watching”

Some participants referred to data as ‘profits’. P7 thinks social media companies *“have to make profit for the business to keep running and dominate the one over the other.”* P11 similarly reflected *“if they’re looking for these things, it’s because there’s profit possibility.”* P12 stated *“There is a reason all these apps are free.”* Further, P12 argued that it is unethical to collect this information, or *“it should be clearly listed how the information should be used”* at least. These quotes emphasize that participants’ perceptions of the social media apps relates to the business model of the apps.

4.7.3 Participants’ Short-Term Reactions to AI/ML Models were Shaped Based on Their Personal Values and Cost-Benefit Trade Offs. Short-term reactions to the models tended to centralize around if the model impacted them (N=6). When asked if location data was a concern to them, P7 replied *“Definitely...because [they] don’t want others to know about [their] location and where [they] move, where [they] stay, where [they] travel.”* After being asked if they see themselves changing behaviors post model exposure, P7 stated

“The social media, like Instagram, posting pictures, starting a record, or just opening a camera. I’ll be more conscious before even opening the app.”

Due to P7 personal privacy considerations, they *did not* like that Instagram was estimating location based on the image before posting. P10, also uncomfortable with the location being estimated, shared P7’s sentiment. P7 and P10 both stated that in the long term, their behavior did change P10 states that they *“will be posting less identifying information on social media,”* although they *“will still scroll through the timeline.”* However, other participants felt like the models didn’t apply to them (N=15). P3 assumed that their privacy is not important, stating

“At the same point, it’s like I’m not at a military base. You know, I’m not in a foreign country. I’m not having state secrets behind me. So it doesn’t really matter that much to me that they’re collecting all of this.”

Whereas, when asked if they would change their behavior, P5 says no *“because [they] don’t open up [their] camera”* and they have not been posting. P9 offers similar reasoning for not changing their behavior as they have already been carefully and only occasionally

posting. What P5, P3, and P9 demonstrate is that their current habits are *unaffected*. Participants who felt unaffected by the models felt less of a reason to stop using the apps. This was most common amongst those who use the app for leisure (N=3). P9 stated they used social media

“for the entertainment aspect for sure. And this is like the distraction in my life. And I guess I’ll just keep [using Instagram].”

Overall, habitual usage may effect short-term reactions to the apps.

4.7.4 Some Participants Experienced a Slight Decline in Social Media Usage After the Study. While six participants reported short-term change, eight reported long-term change. Notably, out of those six, four reported long-term change, and one reported that they did not change after the interview (P20). This implies that four of the long term behavior changes stemmed from those who originally said they felt the models did not apply to them. P5, one of the participants who originally said they would not change, found that

“in addition to using Instagram less frequently I have been using Reddit more instead. It still has not changed the fact that I do not use TikTok, though I am even less keen on opening TikTok links that friends share with me.”

P19 slowed their usage of Instagram particularly, their *“usual place for more personal photos.”* Whereas, P4 mentioned their actions to *“to reduce permissions for these apps to access personal data”* along with their lower usage frequency of social media. P4, P5, and P19 felt, during the interview, that their behaviors likely would not change, but all of them felt *negatively* about the models. P19 described their feelings about their usage during the interview as

“Sometimes I’m able to stop myself and say, is there anything positive about this? And if there’s not, don’t do it. I can turn that into Instagram, saying, is this information that doesn’t need to be out there? And I am protective and private in that way. Unlike some of the other people I know who just put... I think it’ll make me more cautious.”

P4, on the other hand, understood why their friend’s parents wanted their faces blurred in social media posts, implying that they now understood why the parents wanted to be blurred. However, both P4 and P19 expressed hesitance toward changing behaviors, but both reported that they now understood the caution that others took. During the interview, when asked if they felt their behaviors would change, P4 stated,

“But I’m kind of in this spot right now where I feel like I’ve already been compromised, per se. And I don’t know if it would make too much of a difference if I stop now.”

This indicates that negative feelings from the intervention may have influenced participants to change behaviors in the long term.

4.7.5 Model Outputs Felt Personally Offensive. Participants’ personal interpretations to the outputs from the models influenced their reactions and behaviors toward AI/ML models. P18 angrily expressed distain towards TikTok’s gender estimation,

“Because I’m the parent of a generation of kids here in [location], where the assumption is that you just

don't make, they don't make assumptions about gender, period. ... And I am the [parent] of a non-binary child, and so the [protector] kicks in and really pisses me off when people [misgender them]."

Similarly, P13 did not like TikTok estimating the age of children, claiming that it reminded them of "human trafficking". However, out of the two participants *only P13* expressed a change in their long-term behavior. P13 stated in their post-study survey:

"I immediately went into TikTok and changed my settings to deny access to my camera and mic. I use Instagram much less and am more conscious about what I consider posting."

They passed this information onto their young adult children:

"I'm definitely more cautious and wary of social media apps in general. I've shared what I learned with my young adult [children] too."

Notably, P13 was an avid user of both TikTok and Instagram. P18, on the other hand, did not use TikTok or Instagram and reaffirms the cautiousness toward using Instagram "*which [they] do expect to use in the future.*" Arguably, this difference between participants may be due to their personal experiences of feeling impacted by the model.

5 DISCUSSION

In this work, we explored social media users' perceptions and reactions toward transparency in AI/ML models deployed by social media. Our work is motivated by the limited transparency that social media applications offer to their users regarding their data collection and use practices for AI/ML models. Our findings offer insights into ethical and practical opportunities to improve the transparency of AI/ML models. They also highlight several key issues that need further discussion, including integrating transparency into social media applications for their AI/ML models.

5.1 Summary of Findings

Several of the identified themes address our research questions. We found that participants shared the perception that social media use AI/ML for algorithm personalization and/or demographic identification (RQ1). We also demonstrated reactions from participants, comparing both models and their unique positive and negative comments tied to each (RQ2). Finally, we identified that participants' self-reported behavior changes were influenced by how they used the app and/or personal reactions to the model (RQ3).

RQ1: Participants assumed that AI/ML is used for personalization and recommendation algorithms within social media (Section 4.1.1). In their general discussions around AI/ML, participants rarely discussed vision models. However, when directly exposed to vision models, participants tended to assume that social media would implement these models to *collect demographic information* (Section 4.1.2). While all participants expressed similar assumptions towards the purpose of the AI/ML models, their opinions varied on whether they were bothered by data being processed locally (Section 4.4.1). We noticed a prevalent assumption from participants

that social media companies *need* to be profitable and users' recommendation and personalization algorithms are the source of profit (Section 4.4).

RQ2: Based on the prior assumptions of AI/ML in social media, participants assumed that the models were attempting to collect data even when explicitly told that the data may not be collected at all. After revealing both models to participants, we noticed that participants expressed unique positives and negatives towards each. TikTok was criticized for their model always being active with no indication to the user (Section 4.5.1 & Section 4.3). However, participants noted that the amount of data about the user from TikTok's model was noticeably smaller compared to Instagram (Section 4.5.2). Participants were positive about how Instagram's model was executed (i.e., active only upon selection of a photo) but was critical towards the amount of data within the model's output (Section 4.5.2). Our analysis revealed that TikTok's model aligned more with prior participant assumptions (Section 4.6), but Instagram's outputs confused participants (Section 4.5.2). The model outputs from Instagram contained noticeably fewer demographic concepts, which confused participants as the data's purpose was not clear.

RQ3: After seeing the models, some participants self-reported that they noticed small behavior changes, whereas others continued normal usage (Section 4.7). Most participants stated that knowing about the models *would not* change their behaviors as most participants did not actively use the camera for it to concern them (Section 4.7.3). We noticed that participant usage habits had a major influence on participant reflection (Section 4.2). However, participants who valued their privacy to a higher degree did express that they aimed to change their habits after the interview. Our post-interview survey illustrated that more participants *reported long term change* than in the immediate reactions (Section 4.7.4).

5.2 Integrating AI/ML Transparency in Social Media Applications

A general theme found common across all participants is the lack of transparency of how and when social media apps employ local AI/ML. In the following, we discuss how user expectations did not match app behaviors. Then, we identified three transparency preferences discussed by participants. [Finally, we explore a future research agenda based on our findings.](#)

5.2.1 Expectations. Our study showed that participants did not know that their data was being analyzed and were unaware of the consequences of said analysis. Since users were unaware of the possibility that AI/ML could happen to their local images, they did not realize that simple actions such as selecting an image (and deciding not to post it) or opening the camera through the app (accidentally or on purpose) resulted in their data being analyzed. Participants in our study assumed their personal data was analyzed *only after* an image was uploaded. This misconception is not the fault of the user, given that prior to recent developments [20, 110], using a model on a server was the primary method for image analysis. Moreover, we observed that individuals *do not necessarily understand the consequences* of the models. Participants often expressed confusion about Instagram's model and had trouble connecting the model's output to any familiar functionality or use case. This limited understanding

may cause users to be unable to comprehend the consequences of providing their data. How large businesses should collect data is still a debated topic among researchers [9, 57, 113].

5.2.2 Preferences. Participants clearly articulated their expectations of needing more transparency for how and why AI/ML in social media apps process their data. In particular, we identified three preferences. First, some participants wanted to know more about the employed models in terms of inputs (Section 4.5.1), outputs (Section 4.5.2), and purposes (Section 4.7.1). Second, some participants wanted to know *when* the models are being applied (Section 4.4.3). Third, participants expressed interest in controlling or changing how models process their data (Section 4.4.2).

5.2.3 Research Directions. We propose three possible avenues for future research to improve transparency.

Install-time Transparency. Our study emphasizes the important need to be transparent about *when* data is analyzed, *what* a model's inputs and outputs are, and the *purpose* of a model. At *minimum*, participants thought that social media apps should provide notice about model inputs, outputs, and purposes (Section 4.4.2). A potential research direction could be to integrate AI/ML information into existing install-time transparency solutions (privacy policies, privacy dashboards [58], privacy nutrition labels, etc.). However, we foresee three main challenges with integrating AI/ML transparency into existing frameworks: (1) what modality is best for social media users, (2) portraying AI/ML transparency in an accessible way, and (3) providers deploying AI/ML transparency mechanisms.

Some participants indicated they assumed the information about AI/ML they saw was in the apps' privacy policies [71, 105]. However, users likely do not read privacy policies [95] and those who do must parse inaccessible legal language which may confuse users further [47, 109]. Further, some participants claimed they would not trust any transparency mechanism presented by Instagram or TikTok. This raises an interesting research direction for social media AI/ML transparency: what modality would satisfy users of social media? Due to the trust issues unique to social media, existing solutions may require *more transparency* than currently offered by existing frameworks [58, 115].

To address the second challenge, we propose that future work explore user-friendly FactSheets [4] from the XAI community. FactSheets are designed like a schematic for AI/ML containing details about the deployed model, such as the inputs, frequency of inference, outputs, performance test results, and purposes. They are designed to be iteratively updated with the model and contain details about AI/ML risks such as algorithmic consequences of a wrongful prediction. However, FactSheets provide too much information as they are designed for developers of AI/ML. The amount of information within FactSheets could lead to security risks [44] and would likely confuse users further. Striking a balance between relevant information and accessibility is crucial for this hypothetical framework. To address accessibility we argue that providing users with tools that allow for them to interact with the AI/ML could be beneficial [55] and/or explaining AI/ML decisions using human rationale [55].

Our third challenge is how to encourage social media apps to implement the AI/ML transparency mechanisms. For example, prior

work has shown that users have a positive attitude towards Data Safety Labels [54] *but* developers have voiced concerns and challenges during implementation [62]. AI/ML processes can be *complex and non-deterministic*. Developers may not know how a complex model would react with a given input. For example, OpenAI did not foresee lawyers using their models to write legal documents and for their model to hallucinate fake cases [10]. Thus, determining the risks of a complex AI/ML model may not be possible. We need to explore how to properly convey this lack of determinism to users especially as it pertains to their private data.

Run-time Transparency. Participants were interested in an additional level of transparency: being notified at the time when AI/ML is active. In particular, they expressed interest in knowing *when the models were active* to avoid them. Mobile phones offer operating system-level notifications upon access to a camera or microphone. We suggest future works explore designing real-time AI/ML indicators to signify to the user that AI/ML is currently processing their data on the device.

An associated challenge is how much information this notification mechanism can provide the user. One option is to limit the notification mechanism to inform the user of a specific app engaging in AI/ML. Another option is to highlight the type of input data to the model, perhaps using different icons to represent data types. The notification can display additional information in an expandable interface, including the outputs of the model and potentially the purposes.

These notification mechanisms would require collaboration between a mobile platform and the app developers. As opposed to hardware resources, such as the microphone and the camera, AI/ML is an internal app resource, and it is unclear how the operating system can mediate access to the model. As a result, the app has to notify the operating system, and subsequently the user, that AI/ML processing is taking place.

User Control of Model Outputs. As mentioned earlier (Section 4.4.2), some participants felt uncomfortable with the models' observations of their data. Several participants expressed interest in having the option to correct or change the model output. Our participants' preferences present a new challenge to the HCI community: designing user-level controls for local AI/ML models within social media. Such controls would ensure users could still interact with the app without consuming undesirable model outputs. Toward that end, we foresee an interface that enables users to (1) *select* which concepts they want the app to process and (2) change the scores of concepts they find inaccurate.

Providing that level of detail and control comes with many challenges. The first challenge relates to implementation, which, similar to the notification dot, requires collaboration between the mobile platform and the app developer. There are also potential intellectual property issues when revealing model outputs. The second challenge is that models can be complex (e.g., Instagram), providing overwhelming information to a user. Addressing this challenge requires careful design of usable and informative interfaces towards what a user change could mean for their experience. The third challenge, brought by some study participants, is that revealing model outputs would affect users negatively because they would feel the need to optimize their scores. We foresee a complex relationship

between trust and increased transparency. On the one hand, revealing *too much* detail to the user can weaken trust [27]. On the other hand, transparency mechanisms that explain the purposes of potentially intrusive data processing in the context of safety and utility might increase trust.

Finally, some scores *should not be revealed* to users. App functionality may be at risk as complete control over model outputs may interfere with the content a user may see on the app. For example, Instagram may detect concepts like nudity and violence to ensure that they do not show users any inappropriate content. Giving users access to these concepts could cause irreparable harm to users on the platform. For example, if Instagram or TikTok were to use AI/ML processes on local user data to prevent inappropriate content from being uploaded or age verification it would not be beneficial to allow for user control. Future work needs to determine how to discern between AI/ML for application safety and recommendations.

5.3 Limitations

Our work is limited by a set of technical limitations. Both applications included in this study employ techniques to determine if a mobile device is rooted when signing in. Currently, there does not exist a way to circumvent these checks. Due to applications being able to detect when a device is rooted, we **could not** inform participants of the purpose of the collected data. Popular anti-hooking tools like AppDome [3] can detect and alert the application of modern dynamic analysis techniques. Google also offers their Play Store Integrity API [39] to validate if a device is authentic. Rooted devices are not considered authentic and there does not exist a way to bypass this check [18]. Both applications use the Play Integrity API and AppDome-like dynamic analysis prevention tools. The AI/ML tools we explored in this work *were not* protected using these tools, allowing us to build interfaces with them for the participants. However, functions that do utilize these methods downloaded the models, and protected vital components of other areas within the application, limiting our reverse engineering analysis. We acknowledge this limitation and leave this problem for future work as research surrounding Android improves.

A set of factors, other than technical challenges, limits our work. First, we recruited the participants through our institution's large-scale email service, which might have created a regional bias. We also had participants self-report, for example, their own AI/ML familiarity and behavioral changes, which could potentially cause a bias. Therefore, our research does not generalize to all users of social media. Our study population has a much higher representation of educated individuals. We acknowledge that our final sample of participants' has a higher education level overall compared to the educational distribution in social media users in general [67, 94]. However, our sample is consistent with other studies performed in a university setting [2, 32]. Nevertheless, future work is needed to, for example, further investigate how people with less exposure to technology demand transparency and statistically quantify the differences in different populations, based on our observations that they were unaware of the detailed AI model implementations, e.g., how the output is generated.

Another limitation is the specificity of the machine learning models provided in our study. Since we are using real machine learning models deployed by social media applications, we are limited to those models. We only have an opaque understanding of the models themselves and do not know where the data goes after the model gives an output.

6 CONCLUSION

In conclusion, we explored user perceptions and behaviors before and after revealing genuine AI/ML used by Instagram and TikTok. We learned our participants' reactions depended on how much the AI/ML model affected them once learning about it. Participants who posted frequently and/or who interacted with the camera were more likely to react negatively. We also learned that participants felt that *user interactions* were the main source of input for social media AI/ML algorithms. This influenced participant reactions as they were surprised that the camera could analyze them without their knowledge. They were also surprised by the *depth* of the model's output being observed with Instagram's model which lead to several assumptions about the purpose of the data. Our study reported the short-term and long-term behaviors from the participants. We found that short-term behaviors were likely rooted in participant *usage habits*. As for long-term behaviors, we found that 8 out of 21 participants indicated they implemented some long-term behavior change after the study. Our study indicates that the current state of social media algorithm transparency was the main contributor to the confusion harbored by participants. The lack of transparency lead to participants feeling like the AI/ML models were non-consensual, confusing, creepy, and overwhelming. So much so that eight individuals spent less time on the apps after being exposed to the models.

REFERENCES

- [1] National Security Agency. 2024. Ghidra — ghidra-sre.org. <https://ghidra-sre.org/>.
- [2] Nazanin Andalibi and Justin Buss. 2020. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–16. <https://doi.org/10.1145/3313831.3376680>
- [3] AppDome. 2023. How to Prevent Dynamic Hooking of Android Apps — appdome.com. <https://www.appdome.com/how-to/mobile-cheat-prevention/prevent-injection-attacks/prevent-dynamic-hooking-of-android-apps/>.
- [4] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.
- [5] Zahra Ashktorab, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2021. Effects of Communication Directionality and AI Agent Differences in Human-AI Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 238, 15 pages. <https://doi.org/10.1145/3411764.3445256>
- [6] Sumit Asthana, Jane Im, Zhe Chen, and Nikola Banovic. 2024. "I know even if you don't tell me": Understanding Users' Privacy Preferences Regarding AI-based Inferences of Sensitive Information for Personalization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 782, 21 pages. <https://doi.org/10.1145/3613904.3642180>
- [7] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. 2013. "Little brothers watching you": raising awareness of data leaks on smartphones. In *Proceedings of the Ninth Symposium on Usable Privacy and Security* (Newcastle, United Kingdom) (SOUPS '13). Association for Computing

- Machinery, New York, NY, USA, Article 12, 11 pages. <https://doi.org/10.1145/2501604.2501616>
- [8] Jack Bandy and Nicholas Diakopoulos. 2021. More accounts, fewer links: How algorithmic curation impacts media exposure in Twitter timelines. *Proceedings of the ACM on human-computer interaction* 5, CSCW1 (2021), 1–28.
- [9] Anoor Bhimani. 2015. Exploring big data's strategic consequences. *Journal of Information Technology* 30, 1 (2015), 66–69.
- [10] Molly Bohannon. 2023. Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions — forbes.com. <https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/>.
- [11] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.
- [12] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.
- [13] Hongchen Cao, Shuai Li, Yuming Zhou, Ming Fan, Xuejiao Zhao, and Yutian Tang. 2024. Cheating your apps: Black-box adversarial attacks on deep learning apps. *Journal of Software: Evolution and Process* 36, 4 (2024), e2528. <https://doi.org/10.1002/smr.2528> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/smr.2528>
- [14] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable Machine: Approachable Web-Based Tool for Exploring Machine Learning Classification. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '20). Association for Computing Machinery, New York, NY, USA, 1–8. <https://doi.org/10.1145/3334480.3382839>
- [15] Guangyu Chen, Paolo Ciuccarelli, and Sara Colombo. 2022. VisualBubble: Exploring How Reflection-Oriented User Experiences Affect Users' Awareness of Their Exposure to Misinformation on Social Media. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 289, 7 pages. <https://doi.org/10.1145/3491101.3519615>
- [16] Sijing Chen, Lu Xiao, and Akit Kumar. 2023. Spread of misinformation on social media: What contributes to it and how to combat it. *Computers in Human Behavior* 141 (2023), 107643.
- [17] Yi-Shyuan Chiang, Omar Khan, Adam Bates, and Camille Cobb. 2024. More than just informed: The importance of consent facets in smart homes. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 849, 21 pages. <https://doi.org/10.1145/3613904.3642288>
- [18] chiteroman. 2024. GitHub - chiteroman/PlayIntegrityFix: Fix Play Integrity (and SafetyNet) verdicts. — github.com. <https://github.com/chiteroman/PlayIntegrityFix>.
- [19] Kelley Cotter. 2019. Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New media & society* 21, 4 (2019), 895–913.
- [20] Xiangfeng Dai, Irena Spasić, Samuel Chapman, and Bradley Meyer. 2020. The State of the Art in Implementing Machine Learning for Mobile Apps: A Survey. In *2020 SoutheastCon*. IEEE, New York, NY, USA, 1–8. <https://doi.org/10.1109/SoutheastCon44009.2020.9249652>
- [21] Daniel Delmonaco, Samuel Mayworm, Hibby Thach, Josh Guberman, Aurelia Augusta, and Oliver L Haimson. 2024. "What are you doing, TikTok?": How Marginalized Social Media Users Perceive, Theorize, and "Prove" Shadowbanning. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–39.
- [22] Zizhuang Deng, Kai Chen, Guozhu Meng, Xiaodong Zhang, Ke Xu, and Yao Cheng. 2022. Understanding Real-World Threats to Deep Learning Models in Android Apps. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (Los Angeles, CA, USA) (CCS '22). Association for Computing Machinery, New York, NY, USA, 785–799. <https://doi.org/10.1145/3548606.3559388>
- [23] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. How People Form Folk Theories of Social Media Feeds and What it Means for How We Study Self-Presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173694>
- [24] Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms ruin everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 3163–3174. <https://doi.org/10.1145/3025453.3025659>
- [25] Berkeley J. Dietvorst, Berkeley J. Dietvorst, Joseph P. Simmons, Joseph P. Simmons, Cade Massey, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144 (2015). <https://doi.org/10.1037/xge0000033>
- [26] K. Dröge. 2024. noScribe. AI-powered Audio Transcription (Version 0.5)[Computer Software]. <https://github.com/kaixxx/noScribe>.
- [27] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 82, 19 pages. <https://doi.org/10.1145/3411764.3445188>
- [28] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 316, 32 pages. <https://doi.org/10.1145/3613904.3642474>
- [29] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. 2023. Charting the sociotechnical gap in explainable ai: A framework to address the gap in xai. *Proceedings of the ACM on human-computer interaction* 7, CSCW1 (2023), 1–32.
- [30] Abdallah El Ali, Karthikeya Puttur Venkatraj, Sophie Morosoli, Laurens Naudts, Natali Helberger, and Pablo Cesar. 2024. Transparent AI Disclosure Obligations: Who, What, When, Where, Why, How. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 342, 11 pages. <https://doi.org/10.1145/3613905.3650750>
- [31] Ziv Epstein and Hause Lin. 2022. Yourfeed: Towards open science and interoperable systems for social media. *arXiv preprint arXiv:2207.07478* (2022).
- [32] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 2371–2382. <https://doi.org/10.1145/2858036.2858494>
- [33] 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* (Seoul, Republic of Korea) (CHI '15). Association for Computing Machinery, New York, NY, USA, 153–162. <https://doi.org/10.1145/2702123.2702556>
- [34] 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* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3290605.3300724>
- [35] Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2019. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society* 6, 1 (2019), 2053951719860542.
- [36] K. J. Kevin Feng, Xander Koo, Lawrence Tan, Amy Bruckman, David W. McDonald, and Amy X. Zhang. 2024. Mapping the Design Space of Teachable Social Media Feed Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 733, 20 pages. <https://doi.org/10.1145/3613904.3642120>
- [37] Richard Fletcher and Rasmus Kleis Nielsen. 2019. Generalised scepticism: How people navigate news on social media. *Information, communication & society* 22, 12 (2019), 1751–1769.
- [38] Frida. 2024. Frida • A world-class dynamic instrumentation toolkit — frida.re. <https://frida.re/>.
- [39] Google. 2025. Play Integrity API | Google Play | Android Developers — developer.android.com. <https://developer.android.com/google/play/integrity>.
- [40] Google Play Store. 2024. TikTok - Apps on Google Play. https://play.google.com/store/apps/details?id=com.zhiliaoapp.musically&hl=en_US&gl=US.
- [41] Ken Gu, Madeleine Grunde-McLaughlin, Andrew McNutt, Jeffrey Heer, and Tim Althoff. 2024. How Do Data Analysts Respond to AI Assistance? A Wizard-of-Oz Study. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1015, 22 pages. <https://doi.org/10.1145/3613904.3641891>
- [42] Homeboy76. 2022. [Guide] [November 6, 2023] Root Pixel 6 Unlock Bootloader + Pass SafetyNet + Both Slots Bootable + More - xdaforums.com. <https://xdaforums.com/t/guide-root-pixel-6-with-magisk-unlock-bootloader-pass-safetynet-more.4388733/>.
- [43] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. Crowdsourcing the Perception of Machine Teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/>

3313831.3376428

- [44] Yupeng Hu, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li, and Keqin Li. 2021. Artificial intelligence security: Threats and countermeasures. *ACM Computing Surveys (CSUR)* 55, 1 (2021), 1–36.
- [45] Yujin Huang and Chunyang Chen. 2022. Smart App Attack: Hacking Deep Learning Models in Android Apps. *Trans. Info. For. Sec.* 17 (jan 2022), 1827–1840. <https://doi.org/10.1109/TIFS.2022.3172213>
- [46] Yujin Huang, Han Hu, and Chunyang Chen. 2021. Robustness of on-device models: Adversarial attack to deep learning models on android apps. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, IEEE, New York, NY, USA, 101–110.
- [47] Duha Ibdah, Nada Lachtar, Satya Meenakshi Raparathi, and Anys Bacha. 2021. "Why Should I Read the Privacy Policy, I Just Need the Service": A Study on Attitudes and Perceptions Toward Privacy Policies. *IEEE access* 9 (2021), 166465–166487.
- [48] Instagram. 2025. Privacy Policy. <https://privacycenter.instagram.com/policy/>.
- [49] Instagram Help Center. 2024. How Posts Are Chosen for Explore on Instagram. https://help.instagram.com/487224561296752?helpref=faq_content.
- [50] Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 105, 27 pages. <https://doi.org/10.1145/3544548.3581219>
- [51] Bronwyn Jones, Rhianne Jones, and Ewa Luger. 2022. AI 'Everywhere and Nowhere': addressing the AI intelligibility problem in public service journalism. *Digital Journalism* 10, 10 (2022), 1731–1755.
- [52] 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* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. <https://doi.org/10.1145/3491102.3517533>
- [53] Kate Kaye. 2022. Why AI and machine learning are drifting away from the cloud — protocol.com. <https://www.protocol.com/enterprise/ai-machine-learning-cloud-data>.
- [54] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (Mountain View, California, USA) (SOUPS '09). Association for Computing Machinery, New York, NY, USA, Article 4, 12 pages. <https://doi.org/10.1145/1572532.1572538>
- [55] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 250, 17 pages. <https://doi.org/10.1145/3544548.3581001>
- [56] Jan H. Klemmer, Stefan Albert Horstmann, Nikhil Patnaik, Cordelia Ludden, Cordell Burton, Carson Powers, Fabio Massacci, Akond Rahman, Daniel Votipka, Heather Richter Lipford, Awais Rashid, Alena Naiakshina, and Sascha Fahl. 2024. Using AI Assistants in Software Development: A Qualitative Study on Security Practices and Concerns. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security* (Salt Lake City, UT, USA) (CCS '24). Association for Computing Machinery, New York, NY, USA, 2726–2740. <https://doi.org/10.1145/3658644.3690283>
- [57] Panagiota Koukouvinou and Jonny Holmström. 2024. AI Management Beyond Myth and Hype: A Systematic Review and Synthesis of the Literature. *Journal of Academic Research and Trends in Educational Sciences* 17 (2024).
- [58] Hyunsoo Lee, Yugyeong Jung, Hei Yiu Law, Seolyeong Bae, and Uichin Lee. 2024. PrivaAware: Exploring Data Visualization and Dynamic Privacy Control Support for Data Collection in Mobile Sensing Research. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 787, 17 pages. <https://doi.org/10.1145/3613904.3642815>
- [59] Hao-Ping (Hank) Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 775, 19 pages. <https://doi.org/10.1145/3613904.3642116>
- [60] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.
- [61] Jingqi Celeste Li, Morteza Namvar, Ghyiyoung P Im, Saeed Akhlaghpour, et al. 2024. Machine Learning Based Decision-Making: A Sensemaking Perspective. *Australasian Journal of Information Systems* 28 (2024), 315–341.
- [62] Yucheng Li, Deyuan Chen, Tianshi Li, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I. Hong. 2022. Understanding iOS Privacy Nutrition Labels: An Exploratory Large-Scale Analysis of App Store Data. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 356, 7 pages. <https://doi.org/10.1145/3491101.3519739>
- [63] Mengqi Liao and S. Shyam Sundar. 2021. How Should AI Systems Talk to Users when Collecting their Personal Information? Effects of Role Framing and Self-Referencing on Human-AI Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 151, 14 pages. <https://doi.org/10.1145/3411764.3445415>
- [64] Jyoti Mann. 2024. Meta uses your Instagram and Facebook photos to train its AI models — businessinsider.com. <https://www.businessinsider.com/meta-instagram-facebook-photos-used-in-ai-models-training-2024-5?amp>.
- [65] MarketAndMarkets. 2024. AI in Social Media Market Size, Share and Global Market Forecast to 2023 | MarketsandMarkets — marketsandmarkets.com. <https://www.marketsandmarkets.com/Market-Reports/ai-in-social-media-market-92119289.html>.
- [66] Nora McDonald, Sarita Schoenbeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.
- [67] Stacey McLachlan. 2024. Instagram Demographics: Top User Stats for Your Strategy. <https://blog.hootsuite.com/instagram-demographics/>.
- [68] Meta. 2023. DINOv2: State-of-the-art computer vision models with self-supervised learning — ai.meta.com. <https://ai.meta.com/blog/dino-v2-computer-vision-self-supervised-learning/>.
- [69] Meta. 2024. Instagram Feed — transparency.meta.com. <https://transparency.meta.com/features/explaining-ranking/ig-feed/>.
- [70] Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com. <https://ai.meta.com/blog/meta-llama-3/>.
- [71] Meta. 2024. Meta Privacy Policy - How Meta collects and uses user data — privacycenter.instagram.com. https://privacycenter.instagram.com/policy/?section_id=1-WhatInformationDoWe.
- [72] Meta. 2024. Publications - Meta Research — research.facebook.com. <https://research.facebook.com/publications/>.
- [73] MordorIntelligence. 2024. AI in Social Media Market - Forecast Growth & Industry Trends — mordorintelligence.com. <https://www.mordorintelligence.com/industry-reports/ai-market-in-social-media>.
- [74] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
- [75] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- [76] Leila Ouchchy, Allen Coin, and Veljko Dubljević. 2020. AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY* 35 (2020), 927–936.
- [77] European Parliament. 2024. Key Issue 5: Transparency Obligations - EU AI Act — euaiaact.com. <https://www.euaiaact.com/key-issue/5>.
- [78] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. <https://doi.org/10.1145/3411764.3445315>
- [79] Pytorch. 2024. PyTorch — pytorch.org. <https://pytorch.org/mobile/home/>.
- [80] 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* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3173574.3173677>
- [81] Irene Rae. 2024. The Effects of Perceived AI Use On Content Perceptions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 978, 14 pages. <https://doi.org/10.1145/3613904.3642076>
- [82] Yim Register, Lucy Qin, Amanda Baughan, and Emma S. Spiro. 2023. Attached to "The Algorithm": Making Sense of Algorithmic Precarity on Instagram. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 563, 15 pages. <https://doi.org/10.1145/3544548.3581257>
- [83] Daniel Reinhardt, Johannes Borchard, and Jörn Hurtienne. 2021. Visual Interactive Privacy Policy: The Better Choice?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 66, 12 pages. <https://doi.org/10.1145/3411764.3445465>
- [84] Pengcheng Ren, Chaoshun Zuo, Xiaofeng Liu, Wenrui Diao, Qingchuan Zhao, and Shanqing Guo. 2024. DEMISTIFY: Identifying On-device Machine Learning Models Stealing and Reuse Vulnerabilities in Mobile Apps. In *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE)*. IEEE Computer

- Society, IEEE, New York, NY, USA, 468–480.
- [85] Kat Roemmich and Nazanin Andalibi. 2021. Data subjects' conceptualizations of and attitudes toward automatic emotion recognition-enabled wellbeing interventions on social media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.
- [86] Koustuv Saha, Yozen Liu, Nicholas Vincent, Farhan Asif Chowdhury, Leonardo Neves, Neil Shah, and Maarten W. Bos. 2021. AdverTiming Matters: Examining User Ad Consumption for Effective Ad Allocations on Social Media. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 581, 18 pages. <https://doi.org/10.1145/3411764.3445394>
- [87] Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S. B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanatham, and Mauro Cherubini. 2023. Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 505, 23 pages. <https://doi.org/10.1145/3544548.3580848>
- [88] Esha Shandilya and Mingming Fan. 2024. Understanding Older Adults' Perceptions and Challenges in Using AI-enabled Everyday Technologies. In *Proceedings of the Tenth International Symposium of Chinese CHI* (Guangzhou, China and Online, China) (Chinese CHI '22). Association for Computing Machinery, New York, NY, USA, 105–116. <https://doi.org/10.1145/3565698.3565774>
- [89] Tanustree Sharma, Smirity Kaushik, Yaman Yu, Syed Ishtiaque Ahmed, and Yang Wang. 2023. User Perceptions and Experiences of Targeted Ads on Social Media Platforms: Learning from Bangladesh and India. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 663, 15 pages. <https://doi.org/10.1145/3544548.3581498>
- [90] Donghee Shin, Kerk F Kee, and Emily Y Shin. 2022. Algorithm awareness: Why user awareness is critical for personal privacy in the adoption of algorithmic platforms? *International Journal of Information Management* 65 (2022), 102494.
- [91] Erica Shusas and Andrea Forte. 2024. Trust and Transparency: An Exploratory Study on Emerging Adults' Interpretations of Credibility Indicators on Social Media Platforms. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHIEA '24). Association for Computing Machinery, New York, NY, USA, Article 343, 7 pages. <https://doi.org/10.1145/3613905.3650801>
- [92] Christian Sivertsen, Guido Salimbeni, Anders Sundnes Løvlie, Steven David Benford, and Jichen Zhu. 2024. Machine Learning Processes As Sources of Ambiguity: Insights from AI Art. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 165, 14 pages. <https://doi.org/10.1145/3613904.3642855>
- [93] Skylot. 2024. GitHub - skylot/jadx: Dex to Java decompiler – github.com. <https://github.com/skylot/jadx>.
- [94] Statista. 2023. U.S. social network users 2023, by education | Statista. <https://www.statista.com/statistics/1337578/us-distribution-leading-social-media-platforms-by-education/>.
- [95] Nili Steinfeld. 2016. "I agree to the terms and conditions": (How) do users read privacy policies online? An eye-tracking experiment. *Computers in human behavior* 55 (2016), 992–1000.
- [96] Ellie Stevens. 2024. What happens when we train our AI on social media? <https://www.fastcompany.com/91109348/hed-what-happens-when-we-train-our-ai-on-social-media>.
- [97] Google Play Store. 2024. Instagram - Apps on Google Play – play.google.com. https://play.google.com/store/apps/details?id=com.instagram.android&hl=en_US&gl=US.
- [98] František Sudzina. 2017. Do gender and personality traits (BFI-10) influence self-perceived tech savviness. *IT for Practice* 2015 (2017), 195–199.
- [99] Zhichuang Sun, Ruimin Sun, Long Lu, and Alan Mislove. 2021. Mind your weight (s): A large-scale study on insufficient machine learning model protection in the mobile apps. In *30th USENIX Security Symposium (USENIX Security 21)*. 1955–1972.
- [100] Samuel Hardman Taylor and Mina Choi. 2022. An initial conceptualization of algorithm responsiveness: Comparing perceptions of algorithms across social media platforms. *Social Media+ Society* 8, 4 (2022), 20563051221144322.
- [101] Backlinko Team. 2024. Social Network Usage & Growth Statistics (2024) – backlinko.com. <https://backlinko.com/social-media-users>.
- [102] Jane Thier. 2024. AI is moving too fast to keep pace for 4 in 5 workers – finance.yahoo.com. <https://finance.yahoo.com/news/ai-moving-too-fast-keep-190958680.html>.
- [103] TikTok. 2020. How TikTok recommends videos ForYou – newsroom.tiktok.com. <https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you>.
- [104] TikTok. 2024. TikTok's Privacy Innovation 2024 Roadmap | TikTok for Developers – developers.tiktok.com. <https://developers.tiktok.com/blog/2024-privacy-innovation-roadmap>.
- [105] TikTok. 2025. Privacy Policy. <https://www.tiktok.com/legal/page/us/privacy-policy/en>.
- [106] Philip Trammell and Anton Korinek. 2023. *Economic growth under transformative AI*. Technical Report. National Bureau of Economic Research.
- [107] Max Van Kleek, Reuben Binns, Jun Zhao, Adam Slack, Sauyon Lee, Dean Ottewill, and Nigel Shadbolt. 2018. X-Ray Refine: Supporting the Exploration and Refinement of Information Exposure Resulting from Smartphone Apps. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3173574.3173967>
- [108] Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux. 2024. Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 28, 14 pages. <https://doi.org/10.1145/3613904.3642018>
- [109] Kim-Phuong L Vu, Vanessa Chambers, Fredrick P Garcia, Beth Croomkur, John Sulaitis, Deborah Nelson, Russell Pierce, and Robert W Proctor. 2007. How users read and comprehend privacy policies. In *Human Interface and the Management of Information. Interacting in Information Environments: Symposium on Human Interface 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings, Part II*. Springer, Springer, Berlin, Heidelberg, 802–811.
- [110] Ji Wang, Bokai Cao, Philip Yu, Lichao Sun, Weidong Bao, and Xiaomin Zhu. 2018. Deep learning towards mobile applications. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, IEEE, New York, NY, USA, 1385–1393.
- [111] Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefler, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 378, 22 pages. <https://doi.org/10.1145/3613904.3642466>
- [112] Jackson West, Lea Thiemt, Shima Ahmed, Maggie Bartig, Kassem Fawaz, and Suman Banerjee. 2024. A Picture is Worth 500 Labels: A Case Study of Demographic Disparities in Local Machine Learning Models for Instagram and TikTok. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, IEEE, New York, NY, USA, 232–232.
- [113] Martin Wiener, Carol Saunders, and Marco Marabelli. 2020. Big-data business models: A critical literature review and multiperspective research framework. *Journal of Information Technology* 35, 1 (2020), 66–91.
- [114] Lauren Wilcox, Renee Shelby, Rajesh Veeraraghavan, Oliver L. Haimson, Gabriela Cruz Erickson, Michael Turken, and Rebecca Giolotta. 2023. Infrastructuring Care: How Trans and Non-Binary People Meet Health and Well-Being Needs through Technology. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 489, 17 pages. <https://doi.org/10.1145/3544548.3581040>
- [115] Maximiliane Windl, Albrecht Schmidt, and Sebastian S. Feger. 2023. Investigating Tangible Privacy-Preserving Mechanisms for Future Smart Homes. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 70, 16 pages. <https://doi.org/10.1145/3544548.3581167>
- [116] Magdalena Wischnewski, Rebecca Bernemann, Thao Ngo, and Nicole Krämer. 2021. Disagree? You Must be a Bot! How Beliefs Shape Twitter Profile Perceptions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 160, 11 pages. <https://doi.org/10.1145/3411764.3445109>
- [117] Nur Yildirim, Susanna Zlotnikov, Deniz Sayar, Jeremy M. Kahn, Leigh A Bukowski, Sher Shah Amin, Kathryn A. Riman, Billie S. Davis, John S. Minturn, Andrew J. King, Dan Ricketts, Lu Tang, Venkatesh Sivaraman, Adam Perer, Sarah M. Preum, James McCann, and John Zimmerman. 2024. Sketching AI Concepts with Capabilities and Examples: AI Innovation in the Intensive Care Unit. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 451, 18 pages. <https://doi.org/10.1145/3613904.3641896>
- [118] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3290605.3300509>
- [119] Chien Wen (Tina) Yuan, Nanyi Bi, Ya-Fang Lin, and Yuen-Hsien Tseng. 2023. Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 248, 15 pages. <https://doi.org/10.1145/3544548.3580945>
- [120] Rachael Zehrung, Astha Singhal, Michael Correll, and Leilani Battle. 2021. Vis Ex Machina: An Analysis of Trust in Human versus Algorithmically Generated

Visualization Recommendations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 602, 12 pages. <https://doi.org/10.1145/3411764.3445195>

- [121] Shikun Zhang, Yuanyuan Feng, Lujia Bauer, Lorrie Faith Cranor, Anupam Das, and Norman Sadeh. 2021. "Did you know this camera tracks your mood?": Understanding Privacy Expectations and Preferences in the Age of Video Analytics. *Proceedings on Privacy Enhancing Technologies* 2021, 2 (2021), 282--304.
- [122] Xuan Zhao, Cliff Lampe, and Nicole B. Ellison. 2016. The Social Media Ecology: User Perceptions, Strategies and Challenges. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 89-100. <https://doi.org/10.1145/2858036.2858333>

7 APPENDIX

Theme	Sub-theme	Codes	Definition	Example Sub-Codes	
Assumption	ML	Adoption	Participants' adoption of ML models in their lives.	Use for business; Used a lot	
		Capabilities	Participants' original assumption of ML model capabilities.	Data analysis; Provide advertisements	
		Concern	Participants' concern towards ML models at the beginning of the interview.	Data analysis; Provide advertisements	
		Non-Concern	Participants were not concerned about ML models at the beginning of the interview.	Science task; Likes recommendations	
		Confusion	Participants were confused about what ML is or how ML works.	Unsure how ML is used for advertising; Confused about what pictures are real or fake	
		Exposure	Ways of the participants learning about ML models.	Work; None	
		Data Collection	The kind of data participants are considering to be collected by ML models.	Demographics; Interests	
	Instagram	Model	Participants' assumption of IG model's usage.	Human classification; Location estimation	
		Data Usage	Participant's assumption of how IG uses the information collected from the model.	Saved to database; Auto tag images	
		Data Collection	The kind of data participants are assuming to be collected by the IG model.	Ask for location when editing photos; People take photos outside	
		Concern	Concerns participants hold towards IG's data collection behavior they assumed.	Too much data in IG already; Want to be private	
		TikTok	Algorithm	Participants' assumption towards TikTok's algorithm.	Feel to personal; Live updates to recommend better content
			Concern	Participants' concern towards the TikTok model they assumed.	Low accuracy; Would not want anyone estimating gender
			Data Collection	Participants' assumption of how TikTok uses the data being collected.	Sold to advertisers; App development
Model	Participants' assumption on TikTok's ML model.		Accuracy varies; Race can also be assumed with facial features		
	Capabilities	Participants assumption on TikTok's ML model's capability.	Feasible; Does not believe		

Table 2: Themes.

Theme	Sub-theme	Codes	Definition	Example Sub-Codes
Reaction	Social Media	Advertising	Participants' assumption on social media would advertise using collected data.	People get targeted ads; Advertising increased in TikTok
		Algorithm	Participants' assumption on social media's algorithm.	Recommendations come from past usage; Only use data if accuracy is over 99%
		Behavior	Participants' assumed behavior change on their social media usage.	Will be more aware; No change because of rarely posting
		Reflection	Participants' reflection on social media using ML models.	Opinion hinges on data usage; We are forced to use tech as a society
		Concern	Participants' concerns towards social media's ML usage they assumed.	Things appearing to be real; Nefarious data use
		Non-Concern	Participants are not concerned about social media's ML usage they assumed.	Usage Statistics; App improvements only
		Data Collection	Participants assumed items that social media collects.	Actions humans do; Knows faces well
		Data Usage	Participants assumed how social media would use the information they collected.	Necessary for profit; Know who is using the app
	ML	Lesson Reaction	Participants' reaction towards the ML lesson we presented during the interview.	Interesting that it can detect no person; Photos are gathered so quickly
	Instagram	Concern	Participants' concerns towards the IG model we presented.	Granularity is concerning; Model does not meet expectations
		Behavior	Participants' behavior changed after seeing the IG model we presented.	Will still use the app; Will be more aware
		Model	Participants' reaction towards the IG model we presented.	High accuracy; High confidence values seem correct
		Non-Concern	Participants did not express concern towards the IG model we presented.	Comfortable with model; Nothing to hide not concerned
		Reflection	Participants' reflection after seeing the IG model we presented.	Opinions could change; Sobering due to amount of data
		Labels	Participants' reaction on the labels shown on IG model.	So many labels; Seems focused on western ideals
	TikTok	Accuracy	Participants' reaction on how they rate the accuracy of the TikTok model we show.	Terrifying; Gender estimation is correct
		Concern	Participants' concern towards the TikTok model we show.	Age and gender estimation is not generalizable; Different demographics would react differently
		Expectations	Participants' expectations towards TikTok's model.	Lives up to expectations; Technically yes but privacy no
Label		Participants' reaction towards the label on TikTok's model.	Shocked with age; Gender identification is subjective	
Model		Participants' immediate reaction toward the TikTok model we show.	A human would do better; Has a worse opinion of TikTok	
Behavior		Participants' behavior change after seeing TikTok's model we show.	Will not re-download the app; Temporarily change behavior	
Reflection		Participants' reflection on the TikTok model we show.	Positive if they give option; Data collection is similar to colonialism's racism	

Table 3: Themes (continue).

Theme	Sub-theme	Codes	Definition	Example Sub-Codes
Data Collection	Instagram	Concern	Participant's concern towards IG's data collection.	Bad accuracy leads to bad recommendations; Concerned with clothing detection
		Confusion	Participant's confusion on IG's data collection.	Data collection does not seem beneficial; Unaware of uses
		Expectation	Participant's expectation towards IG's data collection.	Data collected would only benefit business
Usage	Social Media	Assumption	Participants' assumption towards social media usage.	Best place for portfolio; Does not see posts from friends and family
		Content Engagement	Participants' content engagement when they use social media.	Does not live stream; Does not use camera
		Posting Habits	Participant's posting habits on social media.	Rarely posts; Posting less when older
		Social Connections	Participants use social media for social connections.	Maintains connections not build an audience; Friends
Privacy and Transparency	Instagram	Concern	Participants' privacy concern towards IG.	Famous people would care more about privacy; Would feel like I was being watched
		Perception	Participants' perception of privacy in IG.	Depends on circumstances; Depends on the person
		Non-Concern	Participants do not have privacy concerns about IG.	Privacy overall is not a concern; Not worried about labels
		Terms of Service	Participants' understanding of Terms of Service in IG.	Wants to be notified of this in terms; Thinks they should be more specific
		Awareness	Participants call for awareness of privacy issues.	People should be aware; Does want to know
	TikTok	Terms of Service	Participants' understanding of Terms of Service in TikTok.	They have to let us know; If its fine print then its user's fault
		Transparency	Participants are not satisfied with the transparency in TikTok's privacy settings.	Children may not be able to consent; Should be transparent
		Concern	Participants' privacy concern towards TikTok.	I'll be a target; Assumed camera frames were not analyzed
	Social Media	Assumption	Participants' assumption on social media privacy terms.	More aware if told; Everyone would disable ML
		Limitations	Participants consider social media's functionality to be limited.	Doesn't know if being able to toggle is possible; Algorithm quality could drop
Non-Concern		Participants do not have privacy concerns about social media.	Purchased items; Consumption patterns	
		Concern	Participants' privacy concern towards social media.	Info can be gleaned from image background; Gore or violence

Table 4: Themes (continue).

Theme	Sub-theme	Codes	Definition	Example Sub-Codes
Deployment	Social Media	Responsibility	Participants consider how to distribute the responsibility of keeping privacy in social media.	Public posts are ok; Do not post to avoid ML
		Terms of Service	Participants' perception of Terms of Service of social media.	SM companies require our data for profit, which is in the terms of service; No one reads them
		Assumption	Participants' assumption on what functionality social media is deploying ML model on.	More useful for them after posting; Only use ML with face filters
		Capabilities	Participants' guessing on what functionality social media is capable of deploying ML model on.	Confident in usage for face filters; Confident with editing photo and face filters
		Confusion	Participants' confusion on how social media could deploy the model in specific functionality.	Too complicated to comprehend; Unsure about opening photo
		Concern	Participants' concern towards these ML model deployments.	Not comfortable with using after posting; Should be better before deploying
		Non-Concern	Participants do not hold concerns towards the ML model deployments.	Fine with all instances; Learning from photo editing is fine
Comparison	Social Media	Expectations	Participants' expectations on what functionality is acceptable to deploy ML models on.	My device = no data for them; Face filter, editing photo, and after posting
		Concern	Participants comparing the concerns they hold for IG and TikTok.	Gender and age feel unsettling; TikTok feels less consensual than IG
		Non-Concern	Participants are comparing and are not concerned about IG or TikTok.	Already assumed this was happening; Data collected is not a problem at all
		Model	Participants are comparing the ML models on IG and TikTok.	Instagram looks for background information; TikTok looks for less labels
		Preference	Participants' preference comparing IG and TikTok ML models.	Instagram; TikTok
		Reflection	Participants' reflection comparing IG and TikTok ML models.	Seeing the models grounded their perspective; Reaction to models not being public

Table 5: Themes (continue).