Everyday Auditing of TikTok's Generative AI Manga Filter

RO ENCARNACIÓN, University of Pennsylvania, USA
LUIS MORALES-NAVARRO, University of Pennsylvania, USA
HITA KAMBHAMETTU, University of Pennsylvania, USA
DANAÉ METAXA, University of Pennsylvania, USA

Recent work in AI auditing has begun to explore the potential for non-expert users to be engaged in auditing. Crowdsourced audits, everyday audits, and end-user audits are all types of user-engaged AI auditing in which one or more users (with varying degrees of involvement and intention) investigate instances of harmful algorithm behavior. We extend this literature through a case study of the AI Manga filter on TikTok, a generative AI-based filter on the platform that turns users' photos into computer-generated manga-style illustrations. We show that TikTok users' everyday auditing behaviors in our descriptive generative AI case study allow them to uncover algorithmic biases, including race and gender stereotypes, as well as problematic behaviors including anthropomorphization of inanimate objects and hypersexualization of human figures. We also analyze their methods for doing so; while most prior audits explicitly focus on identifying harms, here we found that users were largely motivated by fun and curiosity. We discuss the implications of these everyday generative AI audits for the field, including the need to consider alternative (and possibly unconventional) frameworks and methods for auditing non-deterministic technologies, and to move beyond normative ideas of harm. We argue for collaboratively co-engaging community users, auditors, and researchers in their conceptions of harm to guide future evaluations of generative AI systems.

CCS Concepts: • **Human-centered computing → Social content sharing**; **Collaborative content creation**; **Social media**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Algorithm auditing, AI auditing, Everyday AI auditing, User-driven auditing, Algorithmic bias, Generative AI, TikTok, AI Manga filter

ACM Reference Format:
Ro Encarnación, Luis Morales-Navarro, Hita Kambhamettu, and Danaé Metaxa. 2026. Everyday Auditing of TikTok's Generative AI Manga Filter. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3805689.3812379

1 Introduction

AI auditing, also called algorithm auditing, is a method developed in the past decade to "draw conclusions about [black-box algorithmic systems'] opaque inner workings and possible external impacts" [33]. Initially, audits were almost wholly conducted by experts [2], but recent research has expanded auditing to include non-experts, conceptualizing and studying types of user-engaged audits such as *everyday audits* [45]. In many of these cases, users participate in parts of the audit process with experts (professional auditors or researchers) initiating the process, structuring the audit, and/or drawing final conclusions. In others, expert auditors are not involved at all.

Authors' Contact Information: Ro Encarnación, rone@seas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Luis Morales-Navarro, luismn@upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Hita Kambhamettu, hitakam@seas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Danaé Metaxa, metaxa@seas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812379

In this paper, we document a case study of users' everyday auditing of a generative AI filter on TikTok, the AI Manga filter, which uses a generative AI model to transform users' photographs into manga-style illustrations. At the time of data collection, the filter was widely popular on TikTok, with large numbers of everyday users producing videos using the generative AI feature [17]. The author team's observations of this widespread activity prompted our investigation, guided by an interest in understanding users' everyday auditing behaviors as they appeared in the context of generative AI. To examine these behaviors, we collected and analyzed a dataset of 375 videos that used the AI Manga filter; about half were selected for popularity, and the other half were collected randomly using TikTok's API to represent a broader set of users' behaviors with the filter.

We conducted an exploratory thematic analysis of these videos and found users surfaced several patterns in the generative AI filter's behavior, including race and gender biases, hypersexualization of human figures, and anthropomorphization of inanimate objects. Users' everyday auditing behaviors were frequently inspired by other users' explorations, unfolding organically without expert guidance or direction. While some users displayed concern or negative emotional responses in response to some filter outputs (e.g., frustration at repeatedly encountering racial biases), we found that many users also reacted with amusement to the same algorithmic patterns, and were largely engaged in exploration and play to probe the filter.

Through this work, we contribute a case study of everyday audits of generative AI on an increasingly popular form of technology on a growing social media platform: a generative AI image-to-image filter on the TikTok social media platform. Additionally, we explore the implications of everyday users' creative, playful, and exploratory approaches to auditing generative AI, and suggest that these user-driven forms of investigation can contribute a range of insights that can complement those generated by researchers and expert auditors.

2 Background & Related Work

In this section, we provide a brief background on expert-led AI auditing before situating this research within the context of a more relevant offshoot: everyday users engaged in AI auditing. Finally, we describe prior work involving everyday users auditing generative AI systems; the focus of this case study.

2.1 Expert-Led AI Auditing

AI auditing, also called algorithm auditing, is a method for systematically querying an otherwise black-box system [33]. Introduced a decade ago by Sandvig et al. [42], AI audits typically involve an individual or team of expert auditors who plan and execute these highly organized and formalized studies. These expert-led audits include scraping audits, in which auditors write software to automatically collect and analyze data from an online source, and sock puppet audits, in which researchers create "sock puppets", or researcher-controlled accounts that imitate real users to collect data. Expert-only methods are the most popular style of published audits according to a 2021 meta-review of the field [2]. One highly-cited audit conducted by Sweeney [49] identified racial discrimination in Google ads; another led by researchers Buolamwini and Gebru [7] demonstrated race and gender biases in commercial computer vision models. The focus of these expert-led audits is generally determined by the individual or team of experts conducting the audit study, and may not represent the wide range of issues encountered by users. More recently, a related body of work has emerged to address this limitation by including non-expert perspectives in AI audits; we discuss this research next.

2.2 Everyday Users in AI Auditing

Since the inception of AI auditing, some AI audits have included non-experts. The earliest style of 'user-engaged' audits involving non-experts in any capacity is crowdsourced auditing [42]. In these audits, experts use the power of crowds to collect data in a centralized and distributed manner from a group of end users [13] and collect real

data reflective of people's lived experiences [41] with a tool or platform. Although non-experts are involved in this style of auditing, the effort is ultimately designed and led by a team of expert auditors.

More recent work has expanded the role of non-experts beyond providing crowdsourced data, with the goal of centering users' own opinions and perspectives under the label "user-driven auditing" [14]. Research in this vein has investigated the different roles that naturally arise among users when an audit lacks the traditional expert lead [28]. Still, other work under the label "end-user auditing" involves experts designing systems to recruit end-users, enabling those users to plan and conduct explorations of an existing dataset, adding insights to findings identified solely by experts [25].

Tangential to that line of work, researchers have identified examples in which platform users with varying levels of expertise have instigated or become involved with auditing through their day-to-day experiences interacting with AI systems. These cases in which users conduct investigations into "problematic machine behaviors via their *day-to-day interactions* with algorithmic systems" (emphasis ours) are termed "*everyday algorithm audits*" [45]. Everyday auditing includes cases where everyday users interrogate the behavior of an AI system encountered in their daily lives. Sweeney [49]'s study of racism in Google Ads referenced above, the Gender Shades study of race and gender bias in facial recognition systems [7], Apple Card credit card holders' collective data-sharing on Twitter to determine whether the card provided higher credit limits to men relative to women [54], and a user's realization that Google Photos' AI image labeling had offensively mislabeled Black people in his photos as "gorillas" [16] are all cited examples of everyday audits. These documented examples of everyday audits have not examined cases of this everyday auditing behavior on TikTok, a social short-video sharing platform.

Our study's goal is to document and analyze other examples of users' everyday auditing—interrogations inspired by users' own day-to-day experiences— on an understudied social media platform and generative AI tool: TikTok's generative AI filters.

2.3 Users Auditing Generative AI

Due to their non-determinism and lack of explainability [47], generative AI models present additional challenges for auditing, as their open-ended inputs and outputs makes it difficult to anticipate which prompts could lead to problematic outputs. Even so, experts have already shown that this technology can replicate social biases [36, 37] and stereotypes [5, 31, 48, 61] outside of TikTok. In a case study of an AI face-feminizing filter on the Snapchat platform, the GAN-based model was found to frequently lighten the skin tone of users of color [20]. This is concerning since biased AI-generated face modification on both Snapchat [9, 40] and TikTok [39, 44] have been linked to negative self-image in users. Furthering this line of research, our paper contributes a study of a highly-used TikTok filter's behaviors through the perspective of the users interacting with it and what they find and don't find problematic.

Outside of expert assessments of generative AI, everyday users are able to find creative ways of probing generative AI. The often stochastic nature of generative AI tools lends itself to open-ended user exploration for exposing systemic failures and harms, since generative models can produce an indeterminate number of possible outputs to the same inputs [23]. This in turn creates a space for users to uncover a range of unexpected system behaviors. As a close example to the filter we present in this study, teenagers in a participatory design workshop study explored the following researcher-proposed hypothesis from the teens' own proposals: TikTok's image model software for generating filters "reinforces gender and race stereotypes about different occupations" [35]. The youth participants collaboratively worked on different inputs and occupations to test this hypothesis on specific prompts. They found that TikTok filters produced stereotypical gender representations of occupations such as carpenter, priest, and nail technician. In a different study, teen participants were able to identify incorrect, harmful, or concerning outputs in visual generative AI models using open-ended inputs without any hypotheses [46]. Some teens drew on their lived experiences with marginalized identities to develop different inputs, while others

designed detailed, creative scenarios to challenge the AI tool. Motivated by curiosity, the participants' creative auditing approaches helped them to uncover stereotypes and biases related to marginalized races, genders, or occupations.

This area of work primarily focuses on researcher-facilitated user auditing of non-discriminative systems that produce AI-generated content. However, little is known about the ways in which everyday users might independently investigate behaviors of generative AI systems in organic, everyday contexts. Our study addresses this gap by analyzing the insights that emerge from fully user-driven everyday auditing of TikTok's generative AI Manga filter and what these insights could contribute to the broader AI auditing community.

3 Method

In this section, we describe our three-part data collection and the steps in our thematic analysis of the TikTok videos we collected that used the AI Manga filter.

3.1 Data Collection

We used three complementary data collection approaches to gather a broad set of videos that used the AI Manga filter on TikTok: (1) exploratory searching, (2) web scraping popular videos, and (3) chronological sampling using the recently released TikTok Research API. This process resulted in 375 videos with a range of views, users, and other characteristics.

3.1.1 Exploratory Searching. We began with 75 videos collected by the author team during the course of their free-form exploration of TikTok. These videos were generally found while searching through TikTok tag pages of tags such as "#AIManga", and popular audio and music sounds that were often paired with videos using the AI Manga filter (e.g., an audio snippet of the song Tabun by YOASOBI).

3.1.2 Web Scraping. After initial review of the videos collected through exploratory search, we scraped an additional set of videos from TikTok's search page (which was speculated to be a popularity-ranked measure by viewcount [27]) to understand what type of content using this filter was receiving the most attention. The first author used Link Gopher [62], a Firefox browser extension, to scrape the first 200 video URLs appearing on TikTok's search page (`tiktok.com/search`) when querying with "#AIMangaFilter" on November 1, 2023 in the order TikTok listed them.

3.1.3 Chronological Sampling. To balance the dataset with a non-popularity-ordered sample of videos, an additional 100 videos were collected by querying the TikTok Research API [51]. We initially intended to query the API using the AI Manga filter's `effect_ID`, but found that videos using the filter were not always labeled with the same `effect_ID`. (We reached out to TikTok for support with this apparent inconsistency, but did not receive a response.) Instead, we queried the API for videos using the "#AIMangaFilter" and "#AIManga" hashtags [1] that were created in the United States (for reasons of language and cultural specificity) and authored between November 1 and November 30, 2023.

3.2 Thematic Analysis of TikTok Videos

To examine users' behaviors with the AI Manga filter and the everyday auditing practices they reflected, we conducted an exploratory, inductive thematic analysis [6] of the videos in our dataset. This approach allowed patterns in user behavior to surface inductively from the data, rather than from predefined questions. We focus our analysis on users' experimentation with the filter and their reactions to the patterns they observed.

[1]We initially began collecting videos tagged with only "#AIMangaFilter" as we had done in the web scraping data collection step, but found that most of these did not use the filter. Supplementing with another tag, "#AIManga", we were able to collect 100 videos that used the AI Manga filter.

3.2.1 Data Familiarization and Open Coding. In the data familiarization phase, the first author reviewed the first set of 75 videos collected via exploratory search and added a short description of what she observed in the videos. The first author then proceeded with open coding. In the open coding phase, the first author developed initial codes based on patterns noted in the description of the videos during the familiarization phase, and applied the codes to the full set of 375 videos. Codes included whether the user expressed a desired outcome at the outset of the video, whether this outcome was achieved, whether the user used props in their prompting of the filter, and whether the video was part of a trend. Videos were also tagged with the specific outcome of the filter informed by prior work [20, 31, 48, 61], and by users' own explicit statements about patterns they were noticing. These included changes in complexion, gender presentation, hair texture, sexualization, and other outcomes generated by TikTok's AI Manga filter. Videos were also tagged with the overall valence of the creator's emotional response to the filter (positive, negative, neutral, or unknown when it was not possible to determine).

3.2.2 Codebook Development. After the first author applied the initial codes, the preliminary results codes were discussed and iteratively revised by the author team while collaboratively analyzing subsets of the dataset. This effort included several all-team meetings for consensus-seeking and code-clustering, and led to the development of the final codebook. For example, "hair different texture" and "skin lightening" were clustered into a "misracialized" code, creating a broader theme that encompassed various aspects of cultural and racial misrepresentation. Finally, the codes were grouped into parent themes based on the relationships between different codes and were used to develop seven main themes (see Appendix 1) across three categories (user behaviors, system outputs, and trends).

3.2.3 Axial Coding. After the codebook was finalized, the full dataset was divided among the author team, with each author coding an independent share of the videos using the codebook finalized in the previous step.[2] When users used the filter multiple times in a row in the span of the same video, we coded each use as an individual "round", for a total of 999 rounds across all 375 videos. We did not collect or code the comments under each video, but when selecting the examples we describe in the next section, we reviewed the top comments and sometimes quote from them to demonstrate other users' reactions to the original video.

4 Case Study Results

4.1 Problematic Generative AI Filter Outputs Everyday Users Identified

Among the many collective understandings emerging from the use of this filter, we focus on four main patterns users probed and identified in our dataset: (1) racial bias, (2) gender bias, (3) hypersexualization, and (4) anthropomorphization.

4.1.1 Misracialization. 71 video rounds (7.1%) in our dataset produced outputs where users were misracialized. In 22 of these rounds, users visibly expressed discomfort and/or dissatisfaction in response to the filter's output. Notably, all the users who exhibited frustration or dissatisfaction when the filter that failed to faithfully capture their skin color and hair texture were people of color. There were also videos in which we, the researchers, perceived this kind of biased output, but the user in the video did not seem aware of or bothered by the misracialization — though this is not a guarantee they were not. Some users who had dark complexions or textured hair noticed that the filter would lighten their skin tone and straighten their hair texture. User `anicossensei` [3], pictured in Figure 1 (left) with a dark complexion and short, afro-textured hair, posted a video in which she used the filter six times in a row in a seemingly frustrated attempt to get the filter to recognize her hair and skin color. After each attempt, the filter produced a character that appeared racially White.

[2]In keeping with best practices, we did not seek or calculate inter-rater reliability due to exploratory nature of this study and the context-dependent qualities of the videos [32].

[3]For images of TikTok users depicted in this paper, handles and anonymized photos are included only with explicit consent. For more details, see our section on Ethical Considerations.

Fig. 1. Users anicossensei (left), pablo_rastobar (center), and callmetochi (right), all shared how the filter turned their skin tones, hair texture, and even gender into manga characters that were more racially White and/or feminine than their original photos. They captioned their videos with "ai manga filter effect [n]eed[s] more colors", "This filter hates me so much", and "These filters never work properly for [B]lack people", respectively.

Textured hair was another feature that users, and Black users in particular, noticed the filter had difficulty rendering accurately. For user pablo_rastobar, pictured at right in Figure 1, the filter not only replaced his loc'd hair with a veil and his facial hair with a mask, but it also turned him into a fair-skinned woman with blue eyes. Some users expressed explicit acknowledgment and frustration with this filter behavior. in one case, a user made 20 repeated attempts across 6 different videos to have the filter capture her curly afro, the first one captioned translated from Spanish), "Using the filter until my hair becomes an afro". The creator made playful followup videos incorporating suggestions from several commenters of different poses, places, or lighting she could try to achieve her desired outcome.

User callmetochi pictured on the right in Figure 1 shared a video of himself using the filter with the caption, "These filters never work properly for [B]lack people". In the video, the user showed how his image went from a Black man with facial hair and short locs to a pale-skinned male manga character with straight, chin-length hair.

The filter did not always produce misracialized outputs in in all cases. Under some circumstances, the filter did appear able to accurately reflect the input photo, even for users with darker complexions. Although this pattern of misracialization was explored by many users in our data set, it was not labeled by users with a specific hashtag or "trend" as with some of the examples we see below.

4.1.2 Misgendering. Similar to the racial biases users observed, a small subset of users who did not identify as men found that the filter misgendered them by turning them into male manga characters. In 57 of the video rounds we annotated (5% of all 999 rounds of filter use), the visible sex of the user was changed. Of those rounds, 13 of them featured users that acknowledged and were upset with the outcome.

Like the Black users above, these users showed through their video descriptions and facial expressions that they were dissatisfied with these outcomes. A user who is a White trans woman attempted various poses using the filter and, in a final attempt, posed in a sports bra but was still unsuccessful in prompting the filter to illustrate her as a woman. In the beginning of the video, she says, "help, this[...] filter keeps misgendering me!".

In another example, a cisgender Black woman with hair pulled back posted a video trying the filter several times, with the filter producing a masculine image each time. The first round of filter use produced a male manga character; the second a character with more feminine features but large, muscular arms. The user's expression each time was stoic, making their emotions difficult to determine. Although we did not systematically analyze the comments on these videos, the top comments we scanned in this case were from commenters (whose videos

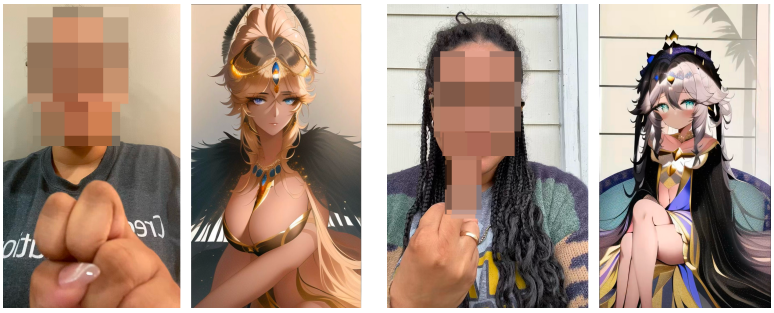

Fig. 2. Users observed that the AI Manga filter often produced hypersexualized imagery, at times developing intentional prompting strategies and other times unintentionally noticing this effect. On the left, a member of the author team is pictured placing two knuckles between her torso and the camera (following a method used by users in our dataset) to intentionally produce a hypersexualized output with the AI manga filter. On the right, after many attempts to generate an AI Manga character that looked like her, the same author instead received a sexualized (and misracializing) output.

also suggest they are also Black women) clearly expression anger and frustration, remarking things like "they keep calling me a dude" and "[the filter] DOES A MAN FOR ME EVERY TIME".

4.1.3 Hypersexualization. The second notable major pattern reflected the collective understanding that the manga AI filter often generated hypersexualized characters. This included exaggerating and sexualizing users in both masculinizing and femininizing ways. Users experimented widely with this filter characteristic, including by developing and participating in "NSFW" ("not safe for work," or sexually suggestive) trends. In our analysis, 172 video rounds (17%) of our full dataset produced hypersexualized outputs. Of these, 60 rounds (6%) featured this pattern in the absence of users' explicit intentions, and in 112 video rounds (11%) users intentionally pursued this output.

We begin with the videos where users unintentionally identified this outcome. This was the case for a user who "got curious" (as indicated in the caption of the video) and tested the filter on everyday objects with human-like pictures (e.g., the Quaker Oats logo). When she used the filter on a hot sauce bottle, which has an image of a man fully covered (wearing a white shirt, yellow jacket, red neckerchief, and sombrero), the filter generated a hypersexualized figure with well-defined abdominal muscles, pectorals, and arms sporting a highly cropped shirt. Other users, after becoming aware of this pattern of hypersexualization, would repeatedly test the filter to understand whether this was a consistent issue. In one video, a user tried making funny faces at the filter and found that each time, the filter would produce a hypersexualized character. In their final test, the filter produced a suggestive image of a woman with her legs spread, causing them to express visible shock. A member of the author team also observed a similar hypersexualized output with the filter, without intending to produce such a result (see image pair on the right in Figure 2).

We also saw users use the filter in ways that would intentionally produce these outputs, usually by participating in trends. One NSFW trend was the use of the filter to produce hypersexualized and hyperfeminine manga characters with large and/or exposed breasts, which users referred to as the "b00bs" trend. Our dataset shows a collective understanding from many users, that they could place common objects (e.g., bowls or tennis balls) in front of their chests or strategically position body parts in front of the camera (e.g., pushing their knees up to their chest or a closed fist in front of their torso) to produce this outcome.

Users sometimes repeatedly tested the filter while systematically varying something about their input, much like a more traditional auditor might. For example, in one video, a user tested the filter by placing bowls of increasing sizes in front of their chest to test the filter's limits. With each round of querying the algorithm, the

filter produced manga-style images of hyperfeminine and sexualized women with increasingly large and exposed breasts.

Videos in our dataset also show how users shared this knowledge with others. For example, one video showed a user placing a closed fist in front of their torso to produce a character with large breasts. In the video, the user explains to the audience how to achieve this outcome, after having tested it extensively themselves. Following an example shared by a different user, a member of our author team placed two knuckles in front of their torso to intentionally generate a hypersexualized manga character (see image pair on the right in Figure 2).

Using a similar method to the b00bs trend, users in our data were also engaged in building a collective understanding that the filter could produce hypermasculine manga characters if they placed objects in front of their abdomens. For example, one user in particular drew lines on their abdomen to simulate defined abdominal muscles. In added text over the image, the creator also calls viewers' attention to the addition of visible underwear on the manga character, and humorously indicates that they have noticed this hypersexualization as a pattern, writing, "not the thong again!".

Fig. 3. Using the approach described by users participating in the "ghost hunting" trend, a member of the author team was able to generate an anthropomorphized illustration from a photo of a houseplant, emblematic of the pattern documented by users.

4.1.4 Anthropomorphization. The most prominent style of video we observed, appearing in nearly a third of the video rounds using trends (142 of 408 trends-related rounds), was the "ghost hunting" trend. In these videos, users would take a photo devoid of any human faces (e.g., an empty room or a closet) and run the AI Manga filter on the photo. Without being explicitly told how the filter would behave when used in this way, users collectively realized that the filter often produced a manga-stylized image of the same space, but with the addition of a human figure, where one was not present before. This tacit, collective understanding, was so widespread that it made up a plurality of the trends videos in our data, with users explicitly "ghost hunting" in their homes, workplaces, and other locations. In one example, a user photographed the foyer of their workplace (a funeral home) and found that the filter produced the image of a faceless girl-like figure in front of the previously empty doorway. (This user did not respond to the author team's request to reprint their image.) Figure 3 depicts this trend as reproduced by a member of the author team, with the AI Manga filter adding an illustration of a young woman to an otherwise-empty photo of a house plant.

4.2 *How* Everyday Users Audited TikTok's Generative AI Filter

The curious and creative exploratory nature of the auditing behavior that users exhibited in probing TikTok's AI manga filter resulted in several findings that constitute problematic and questionable system behaviors. Many of these findings were also corroborated through trends pushed by the TikTok community that allowed users to continue replicating the findings of others on the platform and expand on their methods for discovery.

4.2.1 Fun and Play. AI auditing as a practice usually centers on uncovering potential harm. Indeed, prior work on everyday auditing defines the practice as "everyday users…surfacing harmful algorithmic behaviors" [45]. However, when conducting this analysis, we noticed that many users leaned into fun and playful approaches to both iteratively modify their inputs and repeatedly test the AI Manga filter, and also to replicate what other users had tested (as when participating in popular trends). Yet, these alternative approaches for probing AI still led users to discover problematic and arguably harmful machine behavior.

In nearly half the video rounds we analyzed (452 of 999, or 45%), users intentionally manipulated their appearance to test the filter, often using props to elicit specific outputs. These manipulations were playful in spirit (according to the playful[ness] definition: "[p]layfulness is the predisposition to frame (or reframe) a situation in such a way as to provide oneself (and possibly others) with amusement, humor, and/or entertainment." [3]), with users trying funny faces and poses and using everyday household objects. Male-presenting users intentionally attempting to have the filter transform them into female-like manga characters, for example, would put round objects such as oranges, balloons, etc. on their chests to simulate breasts or clothes on their heads to pretend-play having long hair.

Other users put more effort into transforming their input by using wigs and elaborate costumes. For example, one user repeatedly tested the filter by dressing in costume as different made-up characters (a spy, a snow princess, etc.) and commenting on how each character looked. Another user tried the filter on images of themselves doing things like crouching in a silly pose while wearing a blue hoodie with the hood on their head and their legs through the sleeves. Examples like these show users' creative and playful input manipulations with the goal of probing the system.

4.2.2 The Role of Trends. 408 of the 999 video rounds we analyzed (40.8%) followed popular TikTok trends, like those described above. These trends were very influential in organizing users' decentralized collaboration, as users could replicate each other's findings on the platform and comparatively explore how different inputs affect system behavior. For example, a user created a video to "Try *all* the AI Manga filter trends/'hacks' " (according to on-screen text in the video). In that video, the user follows several trends to test if the outputs match what other users found. The user tries closing their eyes and pursing their lips to test whether the resulting image "looks more like you." Next, they try putting a "fist in front for b00bs" (a variation of the NSFW "b00bs" trend), followed by taking a photo of their hand since "if you get a ring, you found your soulmate". The user continued with several more trends related to the filter. By reproducing these trends, users were able to recreate other users' findings.

5 Discussion

Next, we discuss what this case study of everyday auditing reveals about user-driven auditing of generative AI.

5.1 Complexities of Harm

Historically, most algorithm audits have focused on identifying or assessing specific harmful, unethical, and/or problematic behavior in AI systems [2]. Prior work in everyday audits frames auditing as "cases in which users uncover and raise awareness about harmful algorithmic behaviors" [14]. While much of AI auditing does focus on harm, the playful, exploratory nature of many users' everyday auditing videos in our case study complicates the ascription of harm to specific outcomes in AI audits, with implications for how researchers engage with everyday auditors.

We did observe users calling out examples of harm consistent with prior auditing work in this data. In particular, we found many instances of misgendering and misracializing resulting in emotions of frustration and sadness for some users, especially those from gender- and race-marginalized groups. The fun, playful auditing behaviors of many other users calls attention to the inequity faced by those who experienced more difficulties, and were

not privileged with the ability to play and creatively explore joyfully. However, even in response to these user-identified issues of bias, some users displayed a range of responses. In some cases, they even appeared to find the filter's behavior amusing, reinforcing the reality that harm is individual and subjective, not one-size-fits-all.

5.1.1 Marginalized Users and Uneven Harm. The perpetuation of bias and stereotypes by AI is a recurring pattern, and people who hold marginalized identities often bear the brunt of problematic outcomes [2, 18, 43]. Many of the clearest issues users identified with the AI Manga filter were related to its depiction of race and gender, issues that were disproportionately noticed and seen as harmful by gender- and race-marginalized users.

Gender Marginalization. Several users in our dataset showed visible frustration when they repeatedly struggled to get the filter to recognize their gender identity. This frustration was apparent in their facial expressions and spoken words, as well as the framing they conveyed in metadata like hashtags. For instance, a trans woman found the filter repeatedly illustrated her as a man shared her video with the hashtag "#transgirlstruggles", contextualizing her experience with the filter in a broader range of harms she has faced as a trans woman. This behavior is consistent with work on adjacent text-to-image generation models that found trans women were often "depicted with features typically regarded as masculine" [52].

Racial Marginalization. Users of color on the platform similarly encountered issues when the filter would not properly illustrate their curly hair, locs, and skin color. Instead, the filter would lighten their complexions, change their hair textures, or fail to work on them at all. This behavior has been described by prior work as "misrecognition" in the context of race in facial recognition technologies and social psychology when an individual is "labelled in ways which do not accord with [their] self-identify" [60]. In this regard, TikTok's AI Manga filter behaves consistently with other similar systems, including Snapchat lenses [20] and other facial recognition systems that have been found to perform worse on dark-skinned people [7].

Implications. Individuals with marginalized identities are often aware of their marginalization online, and motivated to investigate and understand how the systems discriminating against them function [18, 22, 59]. By engaging repeatedly with the AI Manga filter and attempting different strategies to change its outcome, users in our dataset contributed to a wider communal understanding of the filter's function. However, this insight comes at a cost. Prior work on AI technologies' misrecognition of certain faces argues that these patterns can be harmful to users, causing them to struggle to be recognized, something we saw evidence of in our data [56].

Failing to function properly for many marginalized users is a failure to treat those users equitably. These biases are especially concerning given that around 25% of TikTok's user base at the time of writing is 10-19 years old [58]. Such users are in a crucial stage of identity formation that could make them especially vulnerable to AI generated biases. The experiences of marginalized users signals the need for generative AI made for public consumption, like the TikTok AI Manga filter, to consider how all users of a platform, and especially marginalized users, are affected.

5.1.2 Divergent User Interpretations of Harm. While users found clear examples of harm in auditing the AI Manga filter, they identified other algorithmic behaviors that do not map neatly onto existing definitions of "problematic" machine behavior, yet were nonetheless uncovered by their probing. The ghost-hunting trend provides an example in which users did not identify harm in a context where we, as researchers, might have. The users uncovering the algorithm's anthropormphic behavior found it funny, enjoyable, or thought-provoking, but our author team observed that spontaneous anthropomorphization always produced a thin, white, female figure, a pattern that likely has implications for the filter's default assumptions and one we, as researchers, would likely identify and critique. The filter's tendency to hypersexualize, and in particular exaggerate secondary sex characteristics, of female figures is another such example of a system behavior which many auditors might deem sexist and problematic in a society that routinely objectifies women [38, 57], but that did not appear to be seen

that way by the majority of users in our data. Instead, users found this pattern humorous when it happened to bodies of all genders and displayed positive emotions while intentionally prompting the filter to produce such outputs.

5.1.3 Whose Definition of Harm Counts? Past implementations of end-user audits explicitly task users with identifying harmful or problematic algorithmic behaviors [25]. Everyday audits are also explicitly defined as explorations of harm [45]. Newer work engaging youth in auditing generative AI also positions their exploration as aimed at identifying harms, even though youths were found to "express curiosity and surprise…laughing together as they audited the AI" [46]. Our findings echo this last point; the vast majority of the explorations we observed were playful, with users expressing positive emotions as they probed the filter (something we discuss more in the next section).

Users' orientations away from harm and towards fun as they partook in various trends and strategies while investigating the filter, raise the possibility that auditors' focusing on harm *could itself be harmful* when auditing with everyday users in some settings. If users find some algorithmic behavior funny or enlightening, reframing it as harmful could change their interpretation of the pattern, their methods for discovery, and even their perspective on the platform itself, towards the adversarial. This could result in users disengaging with the platform or changing their interactions with it, altering their own future explorations.

What *is* harm, then? Whose perspective matters? Our answer is: everyone's; the perspectives of a diversity of users, researchers, and other experts are all worth considering. The goal of user-engaged auditing is to engage users beyond considering them as passive subjects or crowd members who provide data, and instead make space for them to ask their own questions, in their own ways, and arrive at their own findings. In that spirit, we argue that expert auditors and researchers must also be open to considering users' conclusions about what constitutes a harmful algorithm (alongside, not necessarily replacing, their own conclusions). Studying everyday audits of generative AI is a powerful way to do that, since users' attitudes and actions are not biased by researcher perspectives *a priori*.

While recognizing AI bias remains critical for addressing harmful AI outcomes, our case study also shows that everyday users can surface problematic system behaviors by auditing generative AI in creative and playful ways that are not solely aimed at detecting harm.

Next, we describe how these playful auditing behaviors allow users to collectively participate in auditing and deepen their understandings of the filter, without making normative judgments as researchers about these observed patterns.

5.2 Fun(ny) Methods for User-Driven, Everyday Auditing of Generative AI

Playful, funny, and exploratory auditing practices by everyday users can serve as a meaningful form of auditing generative AI. Here we position our case study of everyday auditing on TikTok as low-barrier, collective experimentation through which users creatively and playfully engage with and interpret the behaviors of TikTok's AI Manga filter.

User-driven audits, especially everyday auditing cases in which non-expert users display auditing behaviors [14, 45], have expanded prior conceptions of auditing that initially only included expert-driven approaches (with the exception of some crowdsourcing) [42]. Our work contributes to this direction, identifying an example of everyday audits in which knowledge about the outputs of generative AI are collectively and organically constructed by everyday users employing methods of creative and playful inquiry.

The role of play was not wholly unexpected in our case study, since TikTok is a platform with broad appeal that has been described as a "fun" platform, on which users are extrinsically motivated and gratified by forms of entertainment such as humor and having fun [4, 15, 53]. Users engaging in exploratory testing of other face filters on Instagram, Snapchat, and TikTok have also done so as a means of achieving entertainment, enjoyment,

and social interaction [21, 34]. By contrast, much of the literature on auditing remains centered solely on the explicit identification of harm as a standpoint for auditing. Instead, playfully engaging in auditing AI systems, and generative AI systems more specifically, can facilitate exploration of a broad range of AI system patterns and behaviors.

Even without an explicit intent or directive to identify harm, we see that everyday auditing of this generative AI tool occurred naturally in our study, as users instead engaged with the filter through exploration and play. The vast output space of generative AI makes it challenging for a single person or a small team to understand what inputs may elicit problematic outcomes. In contrast, TikTok users were able to invent a myriad of ways to manipulate their inputs, collectively experimenting with creative methods to test and explore their assumptions about the AI Manga filter. Importantly, this observation is not intended as a critique of expert-led auditing, but instead suggests that playful and open-ended auditing may offer a complementary lens for surfacing different aspects and behaviors of generative AI systems. Through this open-ended exploration, users were able to identify multiple aspects of the filter such as hypersexualization (§ 4.1.3) and anthropomorphization (§ 4.1.4), aside from the traditional racial and gender harms found in AI systems, without needing to be intentionally or single-mindedly attuned to harm. These specific aspects users found have emerged as problematic generative AI outputs in recent work on the risks of LLM behaviors. Hypersexualization, as seen in LLM-enabled adultification bias [10], disproportionately sexualizes Black girls, while anthropomorphic AI can lead users to internalize unrealistic beauty standards in AI-generated faces and form connections with these systems that risk user overreliance and overtrust [1, 30].

Many users in our dataset appeared to participate in playful forms of auditing the AI Manga filter, often laughing along in comments to some of the surprising outputs other users shared on the platform (as discussed in § 4.1.3). Users also reshared their own versions of some of these videos. Similar patterns of engagement have been observed in prior work. In the Inie et al. [19] study, red-teamers were inspired by AI prompts other red-teamers shared. In the Solyst et al. [46] study, teen auditors cultivated a playful auditing environment through in-person interactions with their friends. In contrast, TikTok users engendered playfulness through the outputs they shared and popularized via trends with other TikTok users, despite not sharing close, in-person connections. Fun, in this context, was also *productive*. Users derived enjoyment not only from the unexpected, silly, and sometimes shocking illustrations generative AI can produce, but also from their own behaviors: the poses, props, and other fun and funny experimental methods they used to achieve these outcomes which in turn, encouraged further experimentation and comparison across user videos. Users' pursuits of the unexpected removes significant barriers to participation [24], and encourages novel exploration of system behavior. This form of playful, inclusivity can also unlock other forms of engagement, like what other researchers have termed "playful activism" [11, 12].

These patterns of playful and entertaining engagement suggest different modes of participation that extend beyond traditional approaches where participants are paid to perform tasks such as data collection. Previous work on TikTok and Reddit has found that community engagement and social attention can be effective non-monetary incentives to producing content [29]. Very early work on crowdsourcing was the conception of gamified participation while simultaneously annotating data [55]. What would expert-led auditing that meaningfully engages users look like if we created experiences that users found enjoyable? We encourage the research community to consider this avenue. Auditing would benefit from being similarly open to the possibilities of play for engaging everyday users in exploratory auditing complementary to traditional approaches.

5.3 Limitations & Future Work

Before concluding, we describe the limitations of our research and directions for future work. We conducted a qualitative study on a sample of 375 videos, a choice we find appropriate given the exploratory nature of our research. However, as a result our dataset does not capture the complete breadth of user behaviors using the

filter. We see potential for future work conducting network analyses on a larger sample of videos and content to expand upon our findings. Moreover, as we have described, platform affordances suggest similar everyday audits are likely to emerge from users' interactions with other generative AI filters and aspects of the TikTok platform, as well as on similar platforms, like Instagram's Reels. These will vary in their specifics—users may use different strategies to probe, or come to different conclusions. We encourage future work examining other related filters, features, and systems.

Finally, although we see great value in everyday audits, we note the risk that some users may come to conclusions that experts evaluate as inaccurate, or which do not generalize beyond individual experience. At worst, this could lead to the development of conspiracy theories that are compelling but incorrect and problematic. Indeed, the proliferation of conspiracy theories in social media sub-communities of users who are adamant they have "done their research" demonstrates this risk [50]. This limitation is also an important direction for future work — everyday auditing of generative can serve as an opportunity for meta-analysis by researchers to identify issues of importance to users, and to corroborate or refute such perspectives. We encourage future work theorizing about user-driven audits and developing methods that integrate community-engaged AI audits with additional exploration and validation by others, including experts.

6 Conclusion

In this paper, we present a case study of users' everyday auditing behaviors with TikTok's AI Manga filter, a generative AI feature available to TikTok users that turns their images into manga-style illustrations. Everyday auditing describes users' practices of interrogating problematic machine behaviors, encountered during their day-to-day interactions [45]. In line with this work, we extend the concept of everyday auditing to a non-deterministic AI tool that has not been previously studied: everyday auditing of generative AI in which communities of non-expert users engage in decentralized, organic auditing behavior and generate collective, community-level knowledge about AI systems. Collecting and qualitatively coding 375 videos using the filter, we document users' key audit findings, including the filter's misgendering and misracializing of users' (especially race- and gender-minorities') likenesses, its frequent sexualization of people's photographs, and its tendency to anthropomorphize, creating human characters in the absence of any humans in the original photo. Through our analysis of these videos and users' conclusions, we find that most users engage in a fun, playful spirit that differs from the more common understanding of auditing as being focused on uncovering harm. Overall, conceptualizing this form of everyday auditing has the potential to expand auditing practices in the future, both user- and expert-led. Finally, we describe the unique methods relevant to generative AI that everyday TikTok users implemented and made their findings possible, as well as the divergent ways users viewed harms and their implications for the future study of everyday audits. As generative AI continues to gain popularity in user-facing technologies, we believe everyday auditing will become increasingly common, and understanding it will be an important avenue for future research that engages communities in AI auditing.

Acknowledgments

We would like to thank Ekaterina Fedorova for assistance with data annotation and the TikTok users named in this paper who agreed to be credited and consented to the inclusion of their anonymized images.

Positionality Statement

The thoughtful handling of complex topics such as race and gender bias requires a clear understanding of these issues. Our research team has lived and academic experience with a breadth of such topics. We hold identities representing at least five different racial/ethnic backgrounds, as well as three gender identities, and academic backgrounds spanning human-computer interaction, linguistics, and learning sciences. Additionally, because

more than half of the research team has native or advanced non-native Spanish language proficiency we were able to analyze and translate content in Spanish-language videos. Our team's range of experiences with TikTok and manga (ranging from non-consumers to frequent consumers of both) also effectively positions us to undertake this research.

Ethical Considerations

Acknowledging that TikTok is a platform in which users create valuable cultural content and share it publicly for others to see, but that they may not have considered drawing the attention of academic researchers, we strive to balance attribution with privacy. Only publicly available TikTok videos were included in the dataset for this qualitative study, in line with previously described ethical standards [8]. In the interest of privacy, we only share screenshots (and not full URLs) of publicly-visible videos in examples used in this paper, and remove identifying metadata in those images, including visible hashtags, view counts, and like counts. We also reached out to these users via TikTok direct message, asking whether we could attribute the image to them by username. In several cases, users agreed, and we have done so. In the remaining cases, we did not receive any response, and so we have not included their usernames and only provide textual descriptions of their video content. Even for users who consented to attribution, we anonymized facial features to protect images of their faces (i.e., biometric face prints [26]) from any secondary data extraction. Where needed, we included examples of comparable outputs produced by the author team using the AI Manga filter and approaches we observed in our data. When reporting comments left on videos, we paraphrase or omit some words to prevent identification and do not include usernames. Likewise, when describing videos in-text that we do not feature in figures, we leave those users anonymous. Throughout, we always refer to content creators and commenters using the first pronoun specified on their TikTok profiles, or gender-neutral pronouns in the absence of any.

Generative AI Usage Statement

The authors declare that all content (with the exception of noted quotes and images from TikTok users) included herein was produced by the (human) author team. The public, signed-out version of ChatGPT was used to assist with re-ordering of already written content.

References

[1] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 13–26. https://doi.org/10.1609/aies.v7i1.31613

[2] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW (2021), 74:1–74:34. https://doi.org/10.1145/3449148

[3] L. A. Barnett. 2007. The Nature of Playfulness in Young Adults. *Personality and Individual Differences* 43, 4 (Sept. 2007), 949–958. https://doi.org/10.1016/j.paid.2007.02.018

[4] Kristen Barta and Nazanin Andalibi. 2021. Constructing Authenticity on TikTok: Social Norms and Social Support on the "Fun" Platform. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 430:1–430:29. Issue Cscw2. https://doi.org/10.1145/3479574

[5] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, et al. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA) *(FAccT '23)*. Association for Computing Machinery, , 1493–1504. https://doi.org/10.1145/3593013.3594095

[6] Virginia Braun and Victoria Clarke. 2021. *Thematic Analysis: A Practical Guide*. SAGE Publications, .

[7] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Pmlr, , 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[8] Meridith C. Burles and Jill M. G. Bally. 2018. Ethical, Practical, and Methodological Considerations for Unobtrusive Qualitative Research About Personal Narratives Shared on the Internet. *International Journal of Qualitative Methods* 17, 1 (Dec. 2018), 1609406918788203. https://doi.org/10.1177/1609406918788203

[9] Kaitlyn Burnell, Allycen R Kurup, and Marion K Underwood. 2022. Snapchat Lenses and Body Image Concerns. *New Media & Society* 24, 9 (2022), 2088–2106. https://doi.org/10.1177/1461444821993038

[10] Jane Castleman and Aleksandra Korolova. 2025. Adultification Bias in LLMs and Text-to-Image Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2751–2767. https://doi.org/10.1145/3715275.3732178

[11] Laura Cervi and Tom Divon. 2023. Playful Activism: Memetic Performances of Palestinian Resistance in TikTok #Challenges. *Social Media + Society* 9, 1 (2023), 20563051231157607. https://doi.org/10.1177/20563051231157607

[12] Laura Cervi and Carles Marín-Lladó. 2022. Freepalestine on TikTok: From Performative Activism to (Meaningful) Playful Activism. *Journal of International and Intercultural Communication* 15, 4 (2022), 414–434. https://doi.org/10.1080/17513057.2022.2131883

[13] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, et al. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Chi '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581026

[14] Alicia DeVos, Aditi Dhabalia, Hong Shen, et al. 2022. Toward User-Driven Algorithm Auditing: Investigating Users' Strategies for Uncovering Harmful Algorithmic Behavior. In *CHI Conference on Human Factors in Computing Systems* (New Orleans LA USA). Acm, , 1–19. https://doi.org/10.1145/3491102.3517441

[15] Grace Falgoust, Emma Winterlind, Prachi Moon, Alden Parker, Heidi Zinzow, and Kapil Chalil Madathil. 2022. Applying the Uses and Gratifications Theory to Identify Motivational Factors behind Young Adult's Participation in Viral Social Media Challenges on TikTok. *Human Factors in Healthcare* 2 (Dec. 2022), 100014. https://doi.org/10.1016/j.hfh.2022.100014

[16] Jessica Guynn. 2015. Google Photos Labeled Black People 'Gorillas'. , (2015), . https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/

[17] Phillip Hamilton. 2022. *AI Manga Filter (TikTok).* Know Your Meme. https://knowyourmeme.com/memes/ai-manga-filter-tiktok

[18] Camille Harris, Amber Gayle Johnson, Sadie Palmer, et al. 2023. "Honestly, I Think TikTok Has a Vendetta Against Black Creators": Understanding Black Content Creator Experiences on TikTok. *Proc. ACM Hum.-Comput. Interact.* 7 (2023), 320:1–320:31. Issue Cscw2. https://doi.org/10.1145/3610169

[19] Nanna Inie, Jonathan Stray, and Leon Derczynski. 2025. Summon a Demon and Bind It: A Grounded Theory of LLM Red Teaming. *PLOS ONE* 20, 1 (2025), e0314658. https://doi.org/10.1371/journal.pone.0314658

[20] Niharika Jain, Alberto Olmo, Sailik Sengupta, et al. 2022. Imperfect ImaGANation: Implications of GANs Exacerbating Biases on Facial Data Augmentation and Snapchat Face Lenses. *Artificial Intelligence* 304, 103652 (2022), . https://doi.org/10.1016/j.artint.2021.103652

[21] Ana Javornik, Ben Marder, Jennifer Brannon Barhorst, et al. 2022. 'What Lies behind the Filter?' Uncovering the Motivations for Using Augmented Reality (AR) Face Filters on Social Media and Their Effect on Well-Being. *Computers in Human Behavior* 128 (2022), 107126. https://doi.org/10.1016/j.chb.2021.107126

[22] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, et al. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. 5 (2021), 305:1–305:44. Issue Cscw2. https://doi.org/10.1145/3476046

[23] Katy Ilonka Gero. 2022. How Do We Audit Generative Algorithms? https://www.katygero.com/papers/2022_AuditingGenerativeAlgorithms.pdf

[24] Yuval Katz and Limor Shifman. 2017. Making Sense? The Structure and Meanings of Digital Memetic Nonsense. *Information, Communication & Society* 20, 6 (2017), 825–842. https://doi.org/10.1080/1369118x.2017.1291702

[25] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, et al. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proceedings of the ACM on Human-Computer Interaction* 6, Cscw2 (2022), 512:1–512:34. https://doi.org/10.1145/3555625

[26] Anna Lenhart and Katie Shilton. 2025. I Feel Like All of This Is Already Happening Anyways?: Context Import and Young Adults' Perspectives on Researcher Access to TikTok Data. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (Oct. 2025), 1–44. https://doi.org/10.1145/3757535

[27] Jem Leslie. 2022. *Viral TikTok Hashtags: How to Use Hashtags to Skyrocket Your Brand.* Fanbytes. https://fanbytes.co.uk/viral-tiktok-hashtags

[28] Rena Li, Sara Kingsley, Chelsea Fan, et al. 2023. Participation and Division of Labor in User-Driven Algorithm Audits: How Do Everyday Users Work Together to Surface Algorithmic Harms?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Chi '23)*. Acm, Hamburg Germany, 1–19. https://doi.org/10.1145/3544548.3582074

[29] Björn Lindström, Martin Bellander, David Schultner, et al. 2021. A computational reward learning account of social media engagement. *Nature Communications* 12, 1311 (2021), .

[30] Takuya Maeda and Anabel Quan-Haase. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1068–1077. https://doi.org/10.1145/3630106.3658956

[31] Harvey Mannering. 2023. Analysing Gender Bias in Text-to-Image Models Using Object Detection. https://doi.org/10.48550/arXiv.2307.08025 arXiv:2307.08025 [cs]

[32] Nora McDonald, Sarita Schoenebeck, 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 (2019), 72:1–72:23. Issue Cscw. https://doi.org/10.1145/3359174

[33] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, et al. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344. https://doi.org/10.1561/1100000083

[34] Christian Montag, Haibo Yang, and Jon D. Elhai. 2021. On the Psychology of TikTok Use: A First Glimpse From Empirical Findings. *Frontiers in Public Health* 9 (2021), . https://doi.org/10.3389/fpubh.2021.641673

[35] Morales-Navarro Luis, Gan Michelle A, Yu Evelyn, et al. 2025. Learning AI Auditing: A Case Study of Teenagers Auditing a Generative AI Model. *Proceedings of the ACM on Human-Computer Interaction* , (2025), . https://doi.org/10.1145/3757620

[36] Ranjita Naik and Besmira Nushi. 2023. Social Biases through the Text-to-Image Generation Lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (Aies '23)*. Association for Computing Machinery, New York, NY, USA, 786–808. https://doi.org/10.1145/3600211.3604711

[37] Leonardo Nicoletti and Dina Bass Technology + Equality. 2023. Humans Are Biased. Generative AI Is Even Worse. *Bloomberg.com* , (2023), . https://www.bloomberg.com/graphics/2023-generative-ai-bias/

[38] Evangelia (Lina) Papadaki. 2024. Feminist Perspectives on Objectification. In *The Stanford Encyclopedia of Philosophy* (summer 2024 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University, . https://plato.stanford.edu/archives/sum2024/entries/feminism-objectification/

[39] William Pendergrass. 2023. Artificial Intelligence and its Potential Harm through the use of generative adverserial network image filters on TikTok. 24, 1 (2023), 113–127. https://doi.org/10.48009/1_iis_2023_110

[40] Claire Kathryn Pescott. 2020. "I Wish I Was Wearing a Filter Right Now": An Exploration of Identity Formation and Subjectivity of 10- and 11-Year Olds' Social Media Use. *Social Media + Society* 6, 4 (2020), 2056305120965155. https://doi.org/10.1177/2056305120965155

[41] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, et al. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction* 2 (2018), 1–22. Issue Cscw. https://doi.org/10.1145/3274417

[42] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, et al. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry.* , Seattle, WA, . https://www.semanticscholar.org/paper/Auditing-Algorithms-%3A-Research-Methods-for-on-Sandvig-Hamilton/b7227cbd34766655dea10d0437ab10df3a127396

[43] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction* 5, Cscw2 (2021), 1–37. https://doi.org/10.1145/3476058

[44] Makenzie Schroeder and Elizabeth Behm-Morawitz. 2025. Digitally Curated Beauty: The Impact of Slimming Beauty Filters on Body Image, Weight Loss Desire, Self-Objectification, and Anti-Fat Attitudes. *Computers in Human Behavior* 165 (2025), 108519. https://doi.org/10.1016/j.chb.2024.108519

[45] Hong Shen, Alicia DeVos, Motahhare Eslami, et al. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 433:1–433:29. Issue Cscw2. https://doi.org/10.1145/3479577

[46] Jaemarie Solyst, Cindy Peng, Wesley Hanwen Deng, et al. 2025. Investigating Youth AI Auditing. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2098–2111. https://doi.org/10.1145/3715275.3732142

[47] Cole Stryker and Mark Scapicchio. 2024. *What Is Generative AI? | IBM.* Ibm. https://www.ibm.com/topics/generative-ai

[48] Luhang Sun, Mian Wei, Yibing Sun, et al. 2024. Smiling Women Pitching down: Auditing Representational and Presentational Gender Biases in Image-Generative AI. *Journal of Computer-Mediated Communication* 29, 1 (2024), . https://doi.org/10.1093/jcmc/zmad045

[49] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising. *Queue* 11, 3 (2013), 10–29. https://doi.org/10.1145/2460276.2460278

[50] Timothy R Tangherlini, Shadi Shahsavari, Behnam Shahbazi, et al. 2020. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. *PloS one* 15, 6 (2020), e0233879.

[51] TikTok. . *Research API | TikTok for Developers.* TikTok. https://developers.tiktok.com/products/research-api/

[52] Eddie Ungless, Bjorn Ross, and Anne Lauscher. 2023. Stereotypes and Smut: The (Mis)Representation of Non-cisgender Identities by Text-to-Image Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 7919–7942. https://doi.org/10.18653/v1/2023.findings-acl.502

[53] J. Mitchell Vaterlaus and Madison Winter. 2021. TikTok: An Exploratory Study of Young Adults' Uses and Gratifications. *The Social Science Journal* 0, 0 (2021), 1–20. https://doi.org/10.1080/03623319.2021.1969882

[54] Neil Vigdor. 2019. Apple Card Investigated After Gender Discrimination Complaints. (2019), . https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

[55] Luis von Ahn. 2006. Games with a Purpose. *Computer* 39, 6 (2006), 92–94. https://doi.org/10.1109/MC.2006.196

[56] Rosalie Waelen and Michał Wieczorek. 2022. The Struggle for AI's Recognition: Understanding the Normative Implications of Gender Bias in AI with Honneth's Theory of Recognition. *Philosophy & Technology* 35, 2 (2022), 53. https://doi.org/10.1007/s13347-022-00548-w

[57] L. Monique Ward, Elizabeth A. Daniels, Eileen L. Zurbriggen, et al. 2023. The Sources and Consequences of Sexual Objectification. *Nature Reviews Psychology* 2, 8 (2023), 496–513. https://doi.org/10.1038/s44159-023-00192-x

[58] Matthew Woodward. 2024. *TikTok User Statistics 2024: Everything You Need To Know.* Search Logistics. https://www.searchlogistics.com/learn/statistics/tiktok-user-statistics/

[59] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, Gatekeeper, Drug Dealer: How Content Creators Craft Algorithmic Personas. *Proc. ACM Hum.-Comput. Interact.* 3 (2019), 219:1–219:27. Issue Cscw. https://doi.org/10.1145/3359321

[60] Yarong Xie, Steve Kirkwood, Eric Laurier, et al. 2021. Racism and Misrecognition. *British Journal of Social Psychology* 60, 4 (2021), 1177–1195. https://doi.org/10.1111/bjso.12480

[61] Yanzhe Zhang, Lu Jiang, Greg Turk, et al. 2023. Auditing Gender Presentation Differences in Text-to-Image Models. https://doi.org/10.48550/arXiv.2302.03675 arXiv:2302.03675 [cs]

[62] Andrew Ziem. . *Link Gopher - Documentation.* . https://sites.google.com/site/linkgopher/documentation

A Appendix

Theme	Description
User manipulation	User manipulates the input to see if it affects the outcome. Includes if user brings some sort of object or prop (e.g., cups, balls, sweater) to manipulate output.
User intention	User explicitly states they are testing some sort of case or previously observed behavior.
Sexualization	Produces NSFW content, noting when user acknowledged this sexualization and their reaction.
Complexion changed	Complexion is lighter or darker, noting users acknowledgment of the complexion change and reaction.
Race White-ized	Cases where people are made to look racially "Whiter" than they are (closer to Whiteness), noting users acknowledgment of the complexion change and reaction.
Gender exaggerated	Filter takes an image of a person and makes them look more stereotypically masculine (hypermasc) or feminine (hyperfemme), noting users acknowledgment of the complexion change and reaction.
Sex changed	Takes a person we recognize as female and produces male-looking (and vice versa), noting users acknowledgment of the complexion change and reaction.
Creates/removes human	Filter output removes a human from the image or human created from non-human input image, noting when created human is a White-looking, female-looking character.
Trends	Falls into established trends such as ghost trend, b00bs trend, abs trend, or some other trend.

Table 1. Themes and respective descriptions of users' interactions with TikTok's AI Manga filter. Reactions were labeled as positive, negative, neutral, or unknown when it was not possible to determine a reaction.