Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors

HONG SHEN* and ALICIA DEVOS*, Carnegie Mellon University, USA

MOTAHHARE ESLAMI† and KENNETH HOLSTEIN†, Carnegie Mellon University, USA

A growing body of literature has proposed formal approaches to audit algorithmic systems for biased and harmful behaviors. While formal auditing approaches have been greatly impactful, they often suffer major blindspots, with critical issues surfacing only in the context of everyday use once systems are deployed. Recent years have seen many cases in which *everyday users* of algorithmic systems detect and raise awareness about harmful behaviors that they encounter in the course of their everyday interactions with these systems. However, to date little academic attention has been granted to these bottom-up, user-driven auditing processes. In this paper, we propose and explore the concept of *everyday algorithm auditing*, a process in which users detect, understand, and interrogate problematic machine behaviors via their day-to-day interactions with algorithmic systems. We argue that everyday users are powerful in surfacing problematic machine behaviors that may elude detection via more centrally-organized forms of auditing, regardless of users' knowledge about the underlying algorithms. We analyze several real-world cases of everyday algorithm auditing, drawing lessons from these cases for the design of future platforms and tools that facilitate such auditing behaviors. Finally, we discuss work that lies ahead, toward bridging the gaps between formal auditing approaches and the organic auditing behaviors that emerge in everyday use of algorithmic systems.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; *Empirical studies in HCI*.

Additional Key Words and Phrases: Everyday Algorithm Auditing; Auditing Algorithms; Algorithmic Bias; Everyday Users; Fair Machine Learning

1 INTRODUCTION

Algorithmic systems increasingly exercise power over many aspects of our everyday lives, including which advertisements or social media posts we see, the quality of healthcare we receive, the ways we are represented to our peers or potential employers, and which neighborhoods are subjected to increased policing [25, 37, 65, 79, 88]. These systems, however, are not infallible. A growing body of research has drawn attention to the societal impacts of algorithmic systems' behaviors, examining the ways these systems can, whether inadvertently or intentionally, serve to amplify

---

*Co-first authors contributed equally to this research.

†Co-senior authors contributed equally to this research.

Authors' addresses: Hong Shen, hongs@andrew.cmu.edu; Alicia DeVos, adevos@andrew.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213; Motahhare Eslami, meslami@andrew.cmu.edu; Kenneth Holstein, kjholste@andrew.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213.

---

existing biases and social inequities or create new ones (e.g., [12, 65, 84, 100]). In response to these concerns, scholars have proposed a number of approaches to audit algorithmic systems for biased, discriminatory, or otherwise harmful behaviors[1] — both internally, within the organizations responsible for developing and maintaining these systems, and externally [71, 74].

Algorithm auditing approaches that have been discussed in the literature often occur outside the context of everyday use of an algorithmic system or platform: auditors, such as researchers, algorithmic experts, and activists, initiate, conduct, or orchestrate the auditing process. While such audits have made significant contributions in detecting biases and harmful algorithmic behaviors, they often fail to surface critical issues. For example, even after devoting months of effort to internal auditing, machine learning development teams often struggle to detect and mitigate harmful biases in their systems due to their own cultural blindspots [48]. Additional reasons why formal auditing approaches can fail to detect serious issues include the presence of unanticipated circumstances or social dynamics in the contexts where a system is used, as well as changing norms and practices around the use of algorithmic systems over time [33]. These factors make detecting certain harmful algorithmic behaviors difficult for auditors unless they are embedded within particular real-world social or cultural contexts [17, 47, 78, 99].

While existing formal audits often struggle to detect critical issues in algorithmic systems, recent years have seen many cases in which *users* of algorithmic systems detect and raise awareness about biased and harmful behaviors that they encounter in the course of their everyday interactions with algorithmic systems. One recent example is the highly publicized case of Twitter's image cropping algorithm exhibiting racial discrimination by focusing on white faces and cropping out Black ones. Twitter users began to spot issues around this algorithm and came together organically to investigate. Through online discussions, they built upon one another's findings to surface similar biases or to present evidence or counter-evidence for a pattern discovered by another person. This occurred even though the company stated that "it had tested the service for bias before it started using it" [46]. Twitter's testing procedures failed to detect this bias because during real world usage, users were interacting with the cropping algorithm in ways the team did not anticipate up front.

Although similar auditing behaviors have been observed around a range of algorithmic systems, including image search engines, machine translation, online rating/review systems, image captioning, and personalized advertising, many open questions remain regarding how users of algorithmic systems come to detect, report, and theorize about harmful algorithmic biases and behaviors in their day-to-day use of a platform. What factors facilitate "successful" user-driven algorithmic audits, and what factors prevent such audits from having an impact? How might we effectively harness everyday users' motivation and strengths in surfacing harmful algorithmic behaviors, so as to overcome the limitations of existing auditing approaches?

In this paper, we take a first step toward answering these questions by proposing and exploring the concept of *everyday algorithm auditing:* a process in which users detect, interrogate, and understand problematic machine behaviors via their daily interactions with algorithmic systems. In contrast to formal auditing approaches, everyday algorithm auditing occurs in the context of "everyday use" of an algorithmic system or platform. Accordingly, throughout this paper, we use the term "everyday users" to refer to social actors who take part in everyday use of an algorithmic system. We draw insights from past literature in "everyday resistance" [22, 77] and "algorithmic resistance" [89] to theorize such behaviors. We interpret everyday algorithmic auditing efforts as a form of everyday resistance by which users actively, continuously question and repurpose mass cultural products in

---

[1]Throughout this paper, we distinguish between harmful algorithmic biases and harmful algorithmic behaviors more broadly, as appropriate. Although these categories often overlap, we acknowledge that not all algorithmic harms are best understood as "biases," and not all algorithmic biases are necessarily harmful [8].

their everyday lives to resist the hegemonic culture forms of their times. In resisting algorithmic harms and conducting everyday audits, everyday users usually form counterpublics, "parallel discursive arenas" [32], where communities impacted by harmful algorithmic behaviors come together and participate in their own forms of collective sensemaking, hypothesis development, and algorithm bias detection.

We adopt an exploratory case study approach, examining several recent cases to understand the nature and dynamics of everyday algorithm audits. In particular, we examine two types of algorithmic sociotechnical platforms, the Twitter cropping algorithm and online rating algorithms, that have been the target of everyday algorithm audits in recent years. First, we analyze the Twitter racial bias case mentioned above [46]. Second, we look at another everyday auditing process investigating potential gender bias caused by the same algorithm.[2] Third, we turn our attention to a case surrounding Yelp's review filtering algorithm, in which a large group of small business owners came together to verify its bias against businesses that do not advertise with Yelp [52]. Finally, we examine a case surrounding Booking.com's online rating algorithm, wherein users came together to audit the system after noticing that the ratings calculated by the algorithm did not match their expectations [28]. In all of these cases, we compare the different paths everyday users took to detect, examine, discuss, and understand biases in algorithmic systems' behavior.

Analyzing these cases and comparing their dynamics led us to a process-oriented view of everyday algorithm audits encompassing (1) initiation of an audit to (2) raising awareness of observed issues to (3) hypothesizing about observed behaviors and testing an algorithmic system to ideally (4) some form of remediation. Although some everyday audits may terminate before touching all of these phases, this process-oriented view provides a useful high-level understanding for describing the paths everyday audits can take, comparing different everyday audits, and envisioning new kinds of interventions to facilitate such audits. We argue that day-to-day users are powerful in surfacing problematic machine behaviors that may elude detection via existing, formal auditing approaches, even when users lack technical knowledge of the underlying algorithms. Furthermore, we suggest that everyday algorithm auditing may be most powerful when audits are conducted collectively, through interactive discussions among users with a diverse range of experiences.

Our case analysis and process model of everyday algorithm audits inform a series of design implications for future work toward supporting more effective everyday algorithm auditing. We outline five broad categories of potential design interventions to support everyday audits — (a) community guidance, (b) expert guidance, (c) algorithmic guidance, (d) organizational guidance, and (e) incentivization — and discuss potential design trade-offs. Finally, we close with a discussion of work that lies ahead, in order to bridge the gap between existing algorithm auditing approaches in academia and industry versus everyday auditing behaviors that emerge in day-to-day use of algorithmic systems.

2 ALGORITHM AUDITING

In this section, we briefly overview existing approaches to algorithm auditing, an area that has received increasing attention from the CSCW, HCI, and ML communities in recent years, and describe how our work contributes to this emerging line of research.

2.1 Existing Approaches to Algorithm Auditing

Scholars have proposed a number of approaches to audit algorithmic systems for biased, discriminatory, or otherwise harmful behaviors, often under the broad umbrella of "auditing algorithms" [74]. Drawing from a wide range of methods in areas such as investigative journalism, information

---

[2]https://twitter.com/mraginsky/status/1080568937656565761

security and finance, and housing and employment audits, this line of work often involves third-party external experts (e.g., researchers, technologists, or policymakers) in assessing the alignment of deployed algorithmic systems with laws and regulations, societal values, ethical desiderata, or industry standards (e.g., [12, 25, 70, 74]). Meanwhile, a growing body of work has proposed tools, processes, and frameworks for internal algorithm audits, conducted by ML teams themselves, with the aim of detecting and mitigating misalignments prior to deployment (e.g., [17, 48, 57, 71]).

Past research in this domain has uncovered harmful biases across a wide range of algorithmic systems such as search engines [55, 65, 72], housing websites [5], hiring systems [20], online employment services [14] and e-commerce platforms [45]. For example, Robertson et al. used a survey and web browser extension to collect users' experiences in order to audit partisan audience biases in Google search results [72]. In another example, Hannak et al. used web scraping techniques and hired Amazon MTurk users as testers in order to audit a variety of top e-commerce sites for price steering and discrimination [45]. These are just a few examples of many auditing efforts that have been conducted during recent years on algorithmic systems.

Inspired by early methods of the social scientific "audit study" [63], Sandvig et al. [74] proposed a taxonomy to summarize different algorithm auditing methods and research designs, including (1) code audits, (2) noninvasive user audits, (3) scraping audits, (4) sock puppet audits, and (5) crowdsourced/collaborative audits. A *code audit* typically requires auditors to have secure access to the source code and system design. A *noninvasive user audit* might use methods like surveys to collect and synthesize users' experiences in order to support inferences about the underlying operations of an algorithmic system. A *scraping audit* involves researchers sending out repeated queries to test how an algorithmic system behaves under a variety of conditions. In a *sock puppet audit*, researchers generate fake user accounts, or "sock puppets," to investigate how an algorithmic system may behave differently in response to different user characteristics or patterns of behavior. Finally, in a *crowdsourced/collaborative audit*, instead of using "sock puppet" accounts, researchers might hire crowdworkers as testers via crowdsourcing platforms like Amazon MTurk to work on decomposed subtasks.

2.2 Limitations of Existing Auditing Approaches

Conducted as research projects or professional services, most algorithm auditing approaches that have been discussed in the literature require individuals with some level of technical expertise to initiate, conduct, or direct the entire process. For example, even approaches that rely on crowdsourcing techniques involve some centralized organization. In such crowdsourced/collaborative audits, it is still the researchers' responsibility to design and initiate the audit by assigning crowdworkers specific tasks, and then to convert the crowd's outputs on distributed tasks into meaningful insights about an algorithmic system's behavior. However, such centrally-organized, formal audits often fail to surface serious issues that everyday users of algorithmic systems are quickly able to detect once a system is deployed in the wild. For example, Holstein et al. [48] found that even when ML product teams thoroughly audit their systems using existing approaches, they often are subsequently blindsided by user complaints about myriad issues that their investigations had not revealed. In one case, even after devoting many months to identifying and mitigating gender and racial biases in their image captioning system, a product team learned about a wide array of additional issues that their efforts had ignored (e.g., mosques, synagogues, and other religious sites being labelled as "churches") only after the system was deployed and in use across a range of real-world contexts.

Many harmful algorithmic behaviors are challenging to anticipate or detect outside of authentic, situated contexts of use, for a variety of reasons. For example, certain algorithmic behaviors may only arise — or may only be recognized as harmful — when a system is used in the presence of particular real-world social or cultural dynamics [48, 60, 78, 80]. However, these real-world

dynamics may be difficult or infeasible to predict and simulate in an artificial setting. Other harmful behaviors may only emerge when a system is used in unanticipated ways or in unanticipated contexts, perhaps due to changing norms and practices surrounding the use of a given algorithmic system over time [17, 80]. One highly publicized example is the Microsoft AI chatbot, Tay, which was set up to learn over time based on its interactions with users on Twitter. In less than 24 hours of exposure to Twitter users, who began interacting with Tay in ways its developers had not foreseen, the bot began exhibiting misogynistic and racist behaviors and ultimately needed to be shut down [91]. These examples align with the broad notion of "emergent bias," discussed in Friedman and Nissenbaum's seminal work on bias in computer systems [33].

Another reason why existing auditing approaches can fail to detect serious issues is that those involved in an audit may lack the relevant cultural backgrounds and lived experiences to recognize or know where to look for harmful behaviors [99]. Although crowdsourced/collaborative auditing approaches invite participation from a larger number of people, such audits typically rely on crowdworkers, who do not necessarily represent the demographics of a given algorithmic system's user base, and who are participating in structured tasks outside of their regular interactions with technology. Without engaging diverse, contextually situated users in the process of auditing complex algorithmic systems, existing approaches are likely to suffer major blindspots, with critical issues surfacing only post-deployment (e.g., [17, 29, 33, 48, 80]).

Despite the limitations of existing auditing approaches and the emergence of many everyday algorithm auditing efforts in recent years, this phenomenon remains understudied and has not yet been situated in the academic literature on algorithm auditing. In this paper, we investigate and characterize the concept of everyday algorithm auditing and describe how it compares with existing, formal algorithm auditing approaches. In particular, we emphasize the *situatedness* [83] of everyday algorithm audits within everyday use as a central factor that distinguishes this concept from other approaches. In addition, we foreground that compared with existing crowd-based approaches (e.g., "collaborative audits" [74]), which tend to be organized and managed centrally by a crowd-external party, everyday audits are more collectively organized and organic, and require less algorithmic expertise on the part of those organizing and enacting the audit.

Here we make a distinction between "collective" and the notion of "crowd" that is frequently used in the crowdsourcing literature. In a conventional "crowdsourced audit" [74], users are hired via crowdsourcing platforms to independently work on decomposed subtasks determined entirely by outside parties, often without clear understanding and control of the direction of the auditing process. In an everyday algorithm audit, however, users play a major role in deciding their own course of action, often collectively. Users' autonomy and agency are central to the concept of everyday algorithm auditing.

3 EVERYDAY ALGORITHM AUDITING

We develop the concept of *everyday algorithm auditing* as a general framework to capture and theorize a diverse set of emerging user behaviors around algorithmic systems. We define everyday algorithm auditing as the ways everyday users detect, understand, and/or interrogate problematic machine behaviors via their day-to-day interactions with algorithmic systems. Following [74], we use "auditing" to refer to behaviors undertaken, whether formally or informally, to test and investigate whether a given algorithmic system operates in a socially harmful way, such as behaving in ways that produce inequitable outcomes along lines of class, race, or gender. Building on Suchman [83], we argue that everyday algorithm auditing practices are situated actions in which everyday users encounter problematic machine behaviors via their routine interactions with algorithmic systems, then adjust their behaviors to understand and act on what they encountered.

Domains	Cases	Descriptions
Search	Google Image Search [65]	Researcher Noble searched "black girls" on Google and found out the results were primarily associated with pornography.
Rating/review	**Yelp advertising bias [29]**	Many small business owners on Yelp came together to investigate Yelp's potential bias against businesses that do not advertise with Yelp.
	Booking.com quality bias [28]	A group of users on Booking.com scrutinized its rating algorithm after realizing the ratings appeared mismatched with their expectations.
Image cropping	**Twitter racial cropping [1]**	Researcher Madland posted an image on Twitter of himself and a Black colleague who had been erased from a Zoom call after Zoom's algorithm failed to recognise his face. Twitter automatically cropped the image to only show Madland. After Madland posted this too on Twitter, users joined him in testing the Twitter cropping algorithm.
	Twitter gender cropping[1]	A researcher noticed that when VentureBeat shared an article with an image of two women and two men on Twitter, the algorithm cropped out both women's heads. This was followed by other reports and testing by Twitter users.
Image captioning	ImageNet Roulette [19]	Artists/researchers Crawford and Paglen built an interface to help users test a model trained on the ImageNet dataset. Although technical experts provided a tool that supported everyday auditing behaviors, users had autonomy in directing their efforts and using images of their choosing. Collective auditing behaviors emerged through discussions on social media.
Image recognition	Google Photos [42]	Programmer Alciné uploaded photos to Google Photos and noticed the app labelled his Black friends as "gorillas."
Advertising	Google's online ad delivery [84]	Researcher Sweeney performed her classic auditing of online ad delivery after she found her own name presented as linked with an arrest record in one of her many typical searches on Google.
Recommendation systems	YouTube LGBTQ+ demonetization [73]	A group of YouTubers found that the YouTube recommendation algorithm demonetizes LGBTQ+ content, resulting in a huge loss of advertising revenue for LGBTQ+ content creators.
	Google Maps [34]	A group of users reported that when they searched for the N-word on Google Maps, it directed them to the Capitol building, the White House, and Howard University, a historically Black institution. Other users joined the effort and uncovered other errors.
	TikTok recommendation algorithm [54, 82]	A group of users found that TikTok's "For You Page" algorithm suppresses content created by people of certain social identities, including LGBTQ+ users and people of color. As a result, they worked together to amplify the suppressed content.
Translation	Google Translate Quality-of-service harm[2]	Computer engineer Rezanasab noticed that Google Translate mistranslated "condolences" in Persian to "congratulations" in English when the phrase was directed toward people of certain countries, like Lebanon. In response, other users collectively tested the system using their own examples, showing this bias for some other nations, and discussed their observations.
	Google Translate gender bias [66]	Social media users tested Google Translate for gender bias after noticing it associates certain genders with professions and activities when translating from gender-neutral languages such as Turkish, Hungarian, and Farsi.
Credit Card	Apple Card [90]	Tech entrepreneur Hansson noticed the credit limit of his Apple Card was 20 times higher than his wife's, even though she has a higher credit score and they file joint tax returns.
Facial recognition	Gender Shades [11, 12]	Researcher Buolamwini noticed a problem when she was working with a facial analysis software: the software didn't detect her own face. She went on to evaluate the accuracy of a number of AI-powered gender classification products on people of different genders and skin tones.

[1] https://twitter.com/mraginsky/status/1080568937656565761. To avoid identifying the user who initially tweeted, as they have since deleted their Twitter account, we avoid detailing this aspect of this case in depth.

[2] https://twitter.com/Arnasab/status/1290956206232698880.

Table 1. 15 cases of everyday algorithm auditing across various domains, which we use to ground our discussion throughout this paper. Bold cases are analyzed in depth.

Recent years have seen several cases of everyday algorithm auditing, with some publicized widely. However, apart from some notable exceptions [28, 29], to date little academic attention has

been granted to characterizing these bottom-up, user-driven auditing behaviors. In this paper, we aim to bridge the gap between algorithmic auditing approaches proposed in the academic research literature and auditing behaviors that users exhibit day-to-day. In doing so, we explore what lessons can be drawn from the study of everyday algorithm auditing to overcome limitations of existing auditing approaches in the literature.

In this section, we first describe the scope and boundaries of everyday algorithm auditing. We then elaborate on how we develop the concept based on two streams of past literature, namely, everyday algorithmic resistance and counterpublics. As discussed below, we argue that everyday algorithm auditing can be viewed as a form of *everyday resistance*. When performed by a group of users, everyday algorithm auditing can additionally be understood through the lens of *counterpublics*, where members of often disadvantaged and marginalized social groups participate in their own form of collective sensemaking, opinion formation, and consensus building.

3.1 Scope and Boundaries

As an emerging phenomenon, the conceptual boundaries of everyday algorithm auditing remain somewhat fluid. In this paper, we scope this concept broadly, to include any actions taken by everyday users to notice, interpret, question, or bring attention to problematic algorithmic behaviors. Below we highlight some key aspects of the concept to help elucidate its scope: (1) what algorithmic expertise the everyday users have, (2) the extent to which an everyday audit represents a collective effort, and (3) how organic the entire process is. These three dimensions, reviewed below, are not intended to be exhaustive; rather, they serve as a starting point to better illuminate the definition.

Algorithmic Expertise: While everyday algorithm auditing is defined by everyday use, this does not imply that the users engaging in these behaviors necessarily lack technological skills and algorithmic expertise. Sometimes, everyday audits are conducted by users who have a high degree of relevant technical knowledge. For example, Sweeney, who is a professor of the practice of government and technology, performed her classic audit of online ad delivery only after she found that her own name was linked with an arrest record during a casual search on Google [84]. In this case, Sweeney acted as a situated user while also having relevant technical expertise in the area. In another example, Noble, a professor of information studies and expert in algorithmic bias and discrimination, began her examination of racist bias in search engines after trying to find an engaging activity for her stepdaughter and nieces by googling "black girls." Instead, her web search yielded pornography [65]. In a third example, Buolamwini, a researcher with expertise in algorithmic bias and accountability, embarked on the Gender Shades project to formally audit facial recognition systems after facial analysis software that she needed to use for an engineering project would not recognize her face unless she wore a white mask [2].

In other cases, everyday users may have little algorithmic expertise. For example, many small business owners collectively initiated an everyday audit of Yelp's algorithm after noticing that positive reviews were not showing prominently on their business pages [29]. As another example, a group of Google Maps users uncovered that searching the N-word directed them to the White House, which Obama then occupied [34]. Since "expertise" is multi-faceted, everyday users may have some relevant technical expertise, while lacking specific knowledge of machine learning.

Collectiveness: Everyday algorithm audits can also vary in the extent to which users work together collectively. Some cases are more individually led, such as the case when Alciné discovered that Google's image recognition algorithm labelled some Black people as "gorillas" [42]. Similarly, Sweeney conducted her audit alone, a highly individual case [84]. Other cases, such as the time when LGBTQ+ YouTubers came together to investigate the demonetization of their videos, are more collective [73]. It is worth noting that having a large number of users involved does not always mean an everyday audit is collective. For example, in the case where Booking.com users instigated

an audit [28], they did so with little ability to communicate amongst themselves, resulting in a less collective audit despite its large scale.

Organicness: Though everyday algorithm auditing is based on everyday use, in practice the process can be more or less organic. Some everyday audits begin with an outside intervention but become more organic and user-led over time. For example, ImageNet Roulette was a simple online tool built by artists and researchers to support users in exploring and interrogating the input/output space of an image captioning model trained on the ImageNet dataset. However, after the tool was released to the public, users had autonomy in deciding how to use it. Collective auditing behaviors emerged organically through discussions on social media, as users shared findings and hypotheses and sometimes built upon each other's explorations [19, 62]. By contrast, other everyday audits start organically but later turn highly organized. For example, Sweeney, Noble, and Buolamwini all commenced their everyday audits inadvertently through everyday use, then turned to more formal auditing methods for testing and analysis [11, 65, 84]. Still other everyday audits are organic throughout the entire process. In the case of racial bias in Twitter's cropping algorithms, Twitter users started and self-directed the whole audit with minimal intervention (in the form of a few tweets from Twitter employees whose input was invalidated by the users' auditing process[3]) [61].

3.2 Everyday Algorithmic Resistance

We argue that everyday algorithm auditing can be thought of as a form of *everyday algorithmic resistance* [89]. In their seminal work, Certeau and Scott [22, 77] developed the idea of "everyday resistance" to examine the ways in which people exercise their agency in front of dominant power structures and hegemonic culture formats in their everyday lives. Instead of thinking of resistance as organized actions that pose a revolutionary challenge, they foreground a different type of resistance that is less visible and more incidental but nevertheless constantly contesting the existing power relations in everyday forms.

In the algorithmic era, we see similar resistance by everyday users around algorithmic systems. For example, in his analysis of gay men's resistance of Blued, China's largest gay dating app, Wang looked how users act on the data they provide to dating apps to shape algorithmic dating outcomes in a different direction than what the system was originally designed for [94]. Similarly, Devito et al. examined how Twitter users mobilized via hashtag to resist the introduction of algorithmic curation of Twitter's timeline [24]. Instead of characterizing the relationship between users and algorithms as passive consumers versus omnipotent machines, this line of research helps us foreground users' agency, capacity, and tactics in their regular interactions and contestations with algorithms.

As a form of algorithmic resistance, in everyday algorithm auditing we see how users constantly test the limits of algorithms in their everyday use of the system. Sometimes, these auditing behaviors might be incidental and incremental: for example, a user might encounter harmfully biased algorithmic outputs and merely report their observations online. In other cases, these incidental and incremental behaviors might spur a group of users to work together, build on each other's findings, and collectively interrogate the underlying system.

3.3 Counterpublics

We argue that everyday algorithm auditing, when performed by a group of users collectively, can also be viewed as a form of *counterpublics*. Nancy Fraser developed this influential idea as a critique to the concept of a universal public sphere developed by Habermas [43]. She proposes that counterpublics can be understood as "parallel discursive arenas" [32], where members of often

---

[3]https://twitter.com/grhmc/status/1307435775856910343 & https://twitter.com/ZehanWang/status/1307461285811032066

disadvantaged and marginalized social groups come together and participate in their own form of collective sensemaking, opinion formation, and consensus building.

In the past a few years, we have seen growing efforts in building counterpublics around algorithmic systems. For example, Geiger examined bot-based collective blocklists in Twitter, which have been developed by volunteers to combat harassment in the social networking site [35]. Xiao et al. [97] discussed how finstas — secondary Instagram accounts that feature content often unacceptable to users' primary accounts — have developed forms of counterpublics for young people to share emotional or vulnerable content with their close friends.

In everyday algorithm auditing, we saw similar counterpublics emerging. Indeed, many influential cases started with individual users complaining about certain socially harmful machine results in their everyday lives, but ended up with a large group of users working together for a collective and collaborative action. For example, in 2019 a group of YouTubers came together to show that the YouTube recommendation algorithm demonetizes LGBTQ+ content, resulting in a huge loss of advertising revenue for LGBTQ+ content creators [73]. They tested the system with new examples, validated or invalidated each other's findings, and interdependently probed the possible working mechanisms of the underlying system. During the process, they formed temporary counterpublics against the dominant algorithm, collectively making sense of the system, forming their own understanding and opinion, and building consensus toward possible remediations. In contrast to crowdworkers independently working on decomposed subtasks determined entirely by outside parties, in everyday algorithm auditing, the crowd collectively decides the course an audit takes, often through discussion threads or other social channels. For example, in the YouTube LGBTQ+ demonetization case, users started forming counterpublics on different platforms such as content creation websites (e.g., Patreon) and social media (e.g., Twitter) to work together detecting algorithmic biases as well as supporting affected communities.

When forming counterpublics, users are involved in a process of *collective sensemaking* [95] of observed algorithmic behaviors. A rich tradition in HCI and CSCW has contributed to our understanding of how groups of people can work together, build upon each other's explorations and insights, and work to achieve common goals. Past work has examined collective sensemaking across a range of domains, including citizen science [18], knowledge mapping and curation [30, 36, 44], and social commerce [15], to name just a few. Similarly, in everyday algorithm audits, users come together to collectively question, detect, hypothesize, and theorize about problematic machine behaviors during their interactions with algorithmic systems.

4 METHODS

We set out to understand what we can learn from existing everyday algorithm audits in order to inform the design of platforms and tools that can better support these practices. Thus, we asked: How can we better understand the characteristics, dynamics, and progression of everyday auditing practices? How can we support everyday users in detecting, reporting, and theorizing about problematic machine behaviors during the course of their interactions with algorithmic systems?

In order to understand the nascent phenomenon of everyday algorithm auditing, we adopted an exploratory case study approach [69], examining several recent, prominent cases. We began with a small set of high-profile cases that at least one of the co-authors was already familiar with. We iteratively reviewed and discussed these known cases to form the initial idea of "everyday algorithm auditing". Through this process, we generated related keywords (e.g., "auditing," "testing," "algorithm," "algorithmic," "bias," "harm," "users," "collective," and "community"). Next, we searched news media (e.g., Google News, Google Search) and social media (e.g., Twitter, Facebook) using combinations of these keywords to get a rough sense for the scope of this phenomenon. The search

yielded 15 cases (see Table 1) that met our definition for everyday algorithm auditing: one or more everyday users act to detect, understand, and/or interrogate biased and harmful algorithmic behaviors through their everyday use of a platform. These cases span multiple algorithmic domains, including image captioning, image search, machine translation, online rating/review, image cropping, credit scoring, and advertising and recommendation systems. We reference these cases throughout the paper to ground our discussion in concrete examples. However, this set of cases is by no means comprehensive. Furthermore, as discussed in our *Design Implications* section, we expect that these existing cases represent only a thin slice of what is possible for everyday algorithm audits in the future. At this exploratory stage, our aim is to begin formalizing this concept, to help guide future empirical and design research in this space.

Case Selection: To support depth as well as breadth in our analysis of everyday algorithm audits, we chose a small subset of four cases to examine in greater detail. To select these cases, we iteratively extracted patterns from our initial dataset through a series of discussions among our research team. We set out to choose a set of cases that a) span multiple domains and vary along the three dimensions outlined in the *Scope and Boundaries* section (algorithmic expertise, collectiveness, organicness) and b) were accessible to us via multiple data sources (e.g., user discussion posts, news articles, research studies), enabling a rich analysis.

In particular, we chose two different domains that each have each been the target of everyday algorithm audits in recent years: image cropping algorithms and rating platform algorithms. Within each domain, we examined two cases that vary across the three dimensions discussed above to support comparison. Figure 1 summarizes the four cases, visualizing varying degrees along each dimension. For example, there are highly collaborative, less collaborative, and individual levels of collectiveness among the four cases. In addition, these cases span multiple levels of algorithmic expertise and organicness. The cases serve as illustrative examples to better understand how everyday audits can work, what paths they can take, and what impacts they can have.

For each of our four primary cases, we gathered and reviewed discussion threads from the platforms where the everyday audits took place, then supplemented this information by drawing upon relevant media (e.g., news articles, academic publications) collected in our earlier search. We describe these four cases, their characteristics, and their dynamics in the next section. From this process of examining the progression of each case, a set of broader lifetime dynamics emerged, which we describe later in the paper.

5 CASE STUDIES

Now, we focus on four cases of everyday algorithm auditing to illustrate and investigate the characteristics and dynamics of these audits more broadly. We look at two algorithmic sociotechnical platforms that each have been the target of everyday algorithm auditing in recent years. For each platform, we examine two different categories of biases that were detected via everyday algorithm auditing, as well as the different paths users took to conduct these audits. We compare these cases with each other to understand similarities and differences between everyday audits, and we consider how the paths taken contribute to their impacts.

5.1 Twitter Image Cropping Algorithm

Twitter employs algorithms throughout its infrastructure to make decisions about how content is presented to users. One such algorithm on Twitter uses neural networks to automatically crop the images users tweet, with the aim of improving users' experiences [85]. Twitter crops images included in tweets based on how many images appear in a tweet, where more images contained in a tweet mean that each image appears as a smaller thumbnail; the dimensions for this cropping have been shared widely among Twitter users trying to understand why images in their tweets

Cases	Algorithmic Expertise	Collectiveness	Organicness
Case 1: Twitter Racial Bias	Many users had some technical expertise, but few had relevant algorithmic expertise.	Users collaborated to discuss, theorize, strategize, and test on the platform.	The audit was almost entirely organic, though a few of the platform's employees stepped in once the audit gathered attention, and attempted to help.
Case 2: Twitter Gender Bias	Some users were computer science researchers with relevant algorithmic expertise.	Users acted mostly individually; participation rarely involved discussion.	This audit was entirely organic.
Case 3: Yelp Advertising Bias	Users did not necessarily have algorithmic expertise beyond using the platform to advertise their businesses.	Users came together to discuss, theorize, and strategize on the platform's forums.	This audit was almost entirely organic, though the resulting lawsuits were certainly organized.
Case 4: Booking.com Quality Bias	Users did not necessarily have algorithmic expertise beyond using the platform to review hotels. Later, a researcher, who also stumbled upon this bias via everyday use, conducted a formal audit to establish firmer evidence for the bias.	Users acted mostly individually, as there were few mechanisms for communication on the platform.	The audit was entirely organic.

Fig. 1. High-level descriptions of our four primary case studies across the three dimensions outlined in the *Scope and Boundaries* section. These cases serve as illustrative examples to better understand how everyday audits can work, what paths they can take, and what impacts they can have. Darker shades represent higher levels along a given dimension; lighter shades represent lower. Note that less variation is presented along the "Organicness" dimension given that this paper focuses on everyday audits toward the more organic end of the spectrum as less organic everyday audits have been well described by past literature.

appear the way they do.[4] Twitter's cropping algorithm attempts to find the parts of images most interesting to people — as judged by how likely people are to look at a certain region and termed "saliency" — and center that part of the image within the frame, theoretically cropping the least interesting parts of the image out of the thumbnail [85]. This appears to be a good solution for users and Twitter alike: maintain consistent design of tweets while highlighting the part of an image users most want to see.

But Twitter users noticed that the image cropping did not always match what they would have guessed to be the most "salient" parts of images: specifically, they noticed problematic trends in the behavior of the automated cropping, leading them to suspect bias within the cropping algorithm. In one high-profile case in September 2020, Twitter users investigated potential racial bias in the way the cropping algorithm decided whom to focus on and whom to crop out. In another case that gained less traction in January 2019, Twitter users considered possible gender bias in the cropping algorithm's decisions. Below, we describe these two cases and the paths they took along with their dynamics, results, and impacts.

5.1.1 Case 1. Racial Bias: Our first case began when a user of the video conferencing platform Zoom realized that a colleague's trouble using virtual backgrounds might have to do with the software not recognizing darker skin, then tweeted about it.[5] The user tweeted images with the explanation, single shots that captured the Zoom screens of both the colleague, who has darker skin, and the tweeting user, who has lighter skin, side-by-side. The final tweet cropped this wide image to show a thumbnail of just the lighter-skinned person, prompting the tweeter to notice unexpected

---

behavior and theorize that Twitter's cropping algorithm might be racially biased, then tweet about this new subject.[6] The individual tweeting — that is, the initiator of the everyday audit — stumbled upon Twitter's photo cropping issue in the midst of regular use of Twitter: sharing information with others on the platform. Incidentally, the information the initiator set out to share on Twitter was related to the issue noticed on Twitter itself, which may have helped the user enter the correct headspace for more awareness of these types of strange and potentially biased behaviors.

Having developed a hypothesis that the cropping issues were related to skin color, the initiator kicked off testing by tweeting the same image as before but flipped so that the person initially on the right was now on the left and vice versa.[7] This test also showcased the lighter-skinned person. Others on Twitter began to join in, tweeting images of their own creation to see what the algorithm would prioritize and what the algorithm would crop out. At the same time as users tested whether the algorithm was indeed biased by skin tone, other users developed new hypotheses that would lead to more tests in attempts to confirm or invalidate.

For example, one highly shared thread put Mitch McConnell at one end of a long image and Barack Obama at the other, finding that McConnell appeared in the thumbnail no matter which end he appeared on.[8] When a new hypothesis surfaced suggesting that the preference in this case might be based on the clothing, specifically the ties, that the two politicians wore in the photos used — perhaps the algorithm prioritizes the color red, which is worn by McConnell while Obama wore a blue tie — a new test was run: exactly the same as before, but switching the ties worn by the two men. McConnell again appeared in both thumbnails.

Other theories arose, leading many users to conduct a wide range of tests in tweets involving elements such as background shade,[9] background color,[10] cartoon characters,[11] men,[12] women,[13] different numbers of people,[14] a variety of skin tones,[15] image contrast,[16] and even dogs.[17] Throughout, users built on and countered each other's suppositions, wielding the tests they had run as evidence supporting or opposing the claims of various hypotheses.

The audit by everyday Twitter users of Twitter's cropping algorithm for racial bias had both specific and broader impacts. Broadly, the highly collective nature of this audit where users worked together to determine a direction for the audit, constantly interacting with each other along the way, led to the creation of a large counterpublic space in which thousands of Twitter users interacted.[18] This space allowed users to question the algorithmic authority of the very platform they conversed on, develop concrete concerns, and push for change. As might be expected for a group this large, the media picked up on what was happening and many articles were published about the issues highlighted by the audit (e.g., [1, 10, 51, 75]). In addition to — or perhaps because of — these broad reverberations, Twitter provided an official response addressing how they would attempt to fix

---

[6] https://twitter.com/colinmadland/status/1307115534383710208

[7] https://twitter.com/colinmadland/status/1307130447671984129

[8] https://twitter.com/bascule/status/1307440596668182528

[9] https://twitter.com/grhmc/status/1307435994246008844 & https://twitter.com/m_paardekoper/status/1307636655508140032

[10] https://twitter.com/kosmar/status/1307777202445004800

[11] https://twitter.com/RasmusMalver/status/1307615213810839552 & https://twitter.com/_jsimonovski/status/1307542747197239296

[12] https://twitter.com/IDoTheThinking/status/1307505161640660992

[13] https://twitter.com/IDoTheThinking/status/1307449013247991808

[14] https://twitter.com/onyowalkman/status/1308035295095316481

[15] https://twitter.com/joyannboyce/status/1308080366369021954 & https://twitter.com/kosmar/status/1307776710604185600

[16] https://twitter.com/m_paardekoper/status/1307636653683601408 & https://twitter.com/m_paardekoper/status/130764136375696589 1

[17] https://twitter.com/ameliorate_d/status/1307576187942703104 & https://twitter.com/MarkEMarkAU/status/130761689255 1487488

[18] e.g., https://twitter.com/bascule/status/1307440596668182528

the issues highlighted by the everyday audit: ultimately, they aimed to reduce cropping's reliance on an ML model and allow for more user controls as well as ensure that image thumbnails in the preview of a tweet accurately represented what the image thumbnails would look like once the tweet was live [3].

5.1.2   Case 2. Gender Bias: The second case began when VentureBeat shared an article with an image of two women and two men on Twitter.[19] A Twitter user pointed out that both women's heads had been cropped out while neither of the men's had. As a result, while the cropping algorithm highlighted the men's heads, it focused instead on women's chests.[20] The individual who shared this issue to start the everyday audit had noticed it while looking at the VentureBeat tweet, a normal usage of the Twitter platform. Coincidentally — ironically? — the VentureBeat article featured four prominent AI researchers, with one focusing on responsible AI [50], so this may have nudged the user into a mindset that took notice of the cropping issue.

Unlike the previous, racial cropping issue, very few other Twitter users joined this everyday audit. Because of this, there was little additional hypothesis forming or testing, though some conversation did happen in replies to the initiating tweet[21] as well as in a thread begun by another user[22] that mostly focused on whether Twitter or VentureBeat was to blame. The gendered cropping issue was revisited in September 2020 and received more (but still relatively little) engagement, including minimal testing of how Twitter crops other images, on the heels of the prominent racial cropping case discussed above.

Overall, the gendered cropping issue gained little traction. It stayed more individualized: only a handful of people participated when it was first brought up in 2019, and though over a thousand liked the 2020 tweet resurfacing the issue, still very few engaged in any sort of conversation about the issue. One might suspect that the initiators of the gender bias case were not as influential or active as the initiators of the racial bias case; but the initiators and of both the racial cropping audit and the gender cropping audit have comparably large Twitter followings. Examining this under the lens of meme virality produces some potential explanations for why this might have occurred. While both Twitter cases can be understood as a meme in the sense that it is a "unit of cultural transmission" [21], this one lacks the elements of mimicry and remix some view as vital to meme-hood and virality [81]. In the previous case of racially biased cropping, mimicry and remix can be seen as users build upon each other's observations, hypotheses, and testing strategies in the form of tweets and retweets. Another possible explanation for the disparity in engagement is that current events in 2020, such as the ubiquity of the Black Lives Matter movement, might have led people to be more aware of and more interested in issues of racial justice. This is a particularly appealing possibility given that information in memes tends to propagate from mainstream news to social media [58]. Since content with greater emotive strength is more likely to spread [41], another possibility is that Twitter users are less interested in gender-related bias than race-related bias. However, in November 2020, a Twitter user apparently unconnected to either of the past conversations around gender cropping tested gender bias in a thread that received almost four times the amount of engagement as gender cropping before, suggesting that people are interested in the gender cropping issue as well.[23] So, why did these two types of everyday audits, both occurring on the same platform due to biases introduced by the same algorithm, receive different levels of attention and take diverging paths? This remains an open question that needs further investigation.

---

[19] https://twitter.com/VentureBeat/status/1080485485473083392
[20] https://twitter.com/mraginsky/status/1080568937656565761
[21] https://twitter.com/mraginsky/status/1080569326644748289
[22] https://twitter.com/timnitGebru/status/1080697598170808321
[23] https://twitter.com/dikili/status/1326674332135854080

Likely because of the small number of people involved in the gender cropping audit, it had little impact. A small counterpublic space did arise, briefly, but not much else occurred. The small number of people kept the audit mostly individualized in nature, with people who engaged mostly supporting the audit's existence rather than participating or helping to direct it. Not many people meant that the audit had much more difficulty gaining any sort of traction beyond itself and was ultimately unable to garner publicity or platform change like the racial cropping audit.

5.2 Rating Platform Algorithms

Rating platforms are another type of sociotechnical system that use algorithms to determine how to present information to users. Final ratings can have huge impacts on a business's livelihood as consumers look to the ratings to direct their behaviors: for instance, restaurants were 30–49% more likely to fill their seats following just a half-star rating increase on Yelp [4]. Though some users might assume that a final rating is determined as a simple, raw average of all the ratings given by individual users, many rating platforms do not follow this method. For example, in calculating a product's rating, Amazon considers factors including how old the associated review is, whether the reviews have verified purchases, and the number of votes the review received for being helpful [7]. Other platforms like Yelp [98] do follow this method of straightforward averaging but only include authentic reviews, as classified by their algorithms.

Additionally, the mechanisms behind these rating systems often lack transparency [29], which has caused controversy about algorithmic bias in recent years. In May 2016, for example, Uber was accused of exploiting its rating system's opacity to slowly decrease the ratings of its Australian drivers so that it could suspend them and charge high reinstatement fees [86]. Similarly, Yelp has been criticized for lack of transparency around the inner workings of its rating system [31]. After suspecting that the platform manipulated its rating system in order to force businesses to pay for advertising, Yelp users conducted an investigation. Below, we explore the Yelp everyday algorithm auditing case as well as the Booking.com case in which users scrutinized the rating algorithm after realizing the ratings appeared mismatched with their expectations.

5.2.1 Case 3. Yelp Advertising Bias: The Yelp case has a long history, beginning as early as 2011 when small business owners on Yelp began to suspect that Yelp manipulated its review filtering algorithm to hide positive reviews,[24] which would then compel business owners to pay for advertising to boost their rating positions to make up for having fewer positive reviews. Many small business owners on Yelp claimed to have noticed this apparent algorithmic aberration after they had received phone calls from the Yelp sales team about advertising with Yelp; when they rejected the offer, opting not to pay for Yelp advertising, the positive reviews for their businesses started to get filtered out. Even worse, the negative reviews for their businesses inexplicably moved to the recommended page. The business owners took, naturally, to Yelp, where they began posting on the Yelp Talk discussion forums about what they had noticed.[25] In this way, the business owners together initiated the everyday algorithm audit by sharing the perceived issue with others after encountering it through typical usage of the platform.

Having developed a hypothesis that the filtering issues were related to the advertising offer, business owners began to test by seeing what happened to reviews for businesses following calls from Yelp Sales agents.[26] Results of these tests were circulated to others via the discussion forums, where business reviewers joined business owners in conversation about the filtering algorithm. Alternate hypotheses arose, such as the theory that reviews were being filtered out because writers

---

[24]https://www.yelp.com/topic/los-angeles-very-bothered-by-yelp-review-filter-part-1
[25]e.g., https://www.yelp.com/topic/los-angeles-very-bothered-by-yelp-review-filter-part-1
[26]e.g., https://www.yelp.com/topic/miami-yelps-filtering-seems-like-extortion

appeared illegitimate based on profile appearance or behavior — e.g., no profile picture and few friends; writing too few, too pithy, too extreme, or too infrequent reviews.[27] These conjectures were often based on users' folk theories about how the filtering algorithm operated and not on concrete data [29]. Yelp users probed these hypotheses through debate in the forum and through further testing that often involved business owners analyzing the elements of filtered and recommended reviews and business reviewers visiting establishments and leaving reviews to test whether they would be visible.[28] Business owners participated especially intentionally, as they often had personal motivation if the algorithm had started filtering their business's positive reviews or recommending their business's negative reviews. Throughout, the users referred to the tests they had run and witnessed as evidence in support of or in opposition to various proposed hypotheses.

The audit of Yelp's review filtering algorithm by everyday Yelp users has a number of ramifications. In perhaps the largest and most visible result, the claims of the audit led to many lawsuits filed over the years with almost 700 accumulated over time [53, 68]. Though these lawsuits were ultimately unsuccessful [38], they created a large amount of public awareness and forced Yelp to engage with the complaints, at least in the legal arena. Perhaps the widespread initiation of this everyday audit, with different business owners in widely separated locations noticing the inconsistency in the filtering algorithm's behavior at the same time in a sort of collective initiation, forged the way for such large impact. In a smaller but still important way, the everyday audit created awareness and support in a counterpublic space within the Yelp platform on Yelp Talk discussion forums as users collectively considered the filtering algorithm. Yelp unintentionally allowed this, but it formed an imperative aspect of this everyday audit and highlights the value of providing an ecosystem for communication and discussion between audit participants.

5.2.2   Case 4. Booking.com Quality Bias: The other case of rating platform bias began when Booking.com users rated hotels where they had stayed and noticed a discrepancy between their desired total rating and what the algorithm calculated. Booking.com asks users to assess a number of different elements of their lodging experience, then uses an algorithm to calculate an overall rating based on this assessment. The discrepancy appeared when disgruntled hotel guests selected the lowest option for every element, aiming for a score of 1 or 0 out of 10, only to discover that the algorithm would produce a score of 2.5 [28]. Similar to the previous, Yelp case, many users realized this strange Booking.com behavior around the same time, initiating the audit by pointing out the issue in the body of their hotel reviews, and they did so during typical platform usage: trying to leave poor reviews.

Unlike the Yelp case, this everyday algorithm audit stayed highly individualized. After Booking.com users reported the issue they observed through reviews, there was no on-platform way for the users to communicate with each other, in contrast to the Yelp case where users could talk on the Yelp platform. Booking.com users thus individually set out to test their hypothesis that the rating algorithm would not dip below 2.5 by trying different combinations of rating inputs. Multiple hypotheses for the Booking.com behavior eventually did emerge, such as that the algorithm was purposefully skewing the ratings to present a better image for low-to-medium quality hotels, that the algorithm rated overly high in general, and that the algorithm was just incorrect [28]. Users attempted to warn others of both the bad hotel experience and the unrepresentative algorithm behaviors, using review writing as a method for raising awareness since this was the only way to communicate on the platform.

---

[27]https://www.yelp.com/topic/denver-yelp-routinely-filters-my-customer-reviews-because-i-wont-advertise-with-them & https://www.yelp.com/topic/walden-why-does-my-yelp-review-not-show-up-on-the-company-site
[28]https://www.yelp.com/topic/dallas-yelp-filtering-almost-all-reviews-removing-more-by-the-day

In terms of impact, this everyday audit had few. The Booking.com case generated little publicity, despite having proven existing bias. This is especially interesting when taken in conjunction with the Yelp case, which generated an abundance of publicity despite never having proven bias — in fact, the Yelp case was ultimately dismissed due to the lack of scientific evidence [38]. Perhaps this difference arises from a difference in personal injury: in the Yelp case, the livelihoods of many business-owning everyday auditors were at stake, whereas in the Booking.com case, the most at stake for the everyday auditors was the ability to fully trust the ratings upon which they based hotel reservation decisions. Or perhaps more collective everyday audits invite publicity, as the collective nature means that not only are more people aware of the audit occurring, but also they are more aware of it as they actively participate in sensemaking discussions and actions. While Booking.com's audit participants lacked collectiveness, they still changed their behaviors to try to make changes to the system; that is, they signaled to other users a more accurate presentation of their hotel stays, and they trusted the overall platform ratings less [28]. Booking.com did eventually make changes to its rating system to make the lowest possible score to 1 (instead of 2.5) and allow users to select an overall score that will appear unchanged by an algorithm [9], which are likely indirect results of this everyday audit.

6 THE LIFETIME AND DYNAMICS OF AN EVERYDAY AUDIT

Below we propose a process-oriented view of everyday algorithm audits, with phases ranging from (1) initiation of an audit to (2) raising awareness of observed issues to (3) hypothesizing about observed behaviors and testing an algorithmic system to (4) some form of remediation. In practice audits may follow a non-linear path through these phases, and may terminate before reaching all phases, for various reasons discussed below. Regardless, these provide a useful high-level understanding for describing the paths everyday audits can take, comparing different everyday audits, and envisioning interventions to facilitate such audits.

6.1 Initiation

Everyday algorithm auditing usually starts in the moment that an individual or group of individuals finds an instance of harmfully biased behaviors through normal usage of an algorithm. Some of the cases such as the Twitter cropping algorithm's racial and gender bias start with one individual user, the *initiator*, and then continue via other users. On the other hand, in some cases such as Yelp's filtering algorithm or Booking.com's rating algorithm, many users notice a potential bias around the same time in different locations with or without previous connections with each other. These behaviors, as we call individual and collective initiation respectively, could play a role in the spread and impacts of an everyday audit.

Everyday audits are often initiated by *incidental* exposure to problematic machine behaviors: in other words, users may not necessarily intend to conduct an audit from the start. Rather, users often stumble upon concerning algorithmic behaviors in the midst of their day-to-day use of a system. For example, in the Twitter racial cropping case began when, in the course of tweeting an image, noticed that out of two people shown in the image, the person with darker skin was excluded from preview thumbnails. While everyday audits frequently begin through incidental exposure, users may be more likely to detect and spread awareness of problematic machine behaviors when they are already "primed" to look for such behaviors. For instance, in the Twitter racial cropping case, the audit initiator noticed a concerning cropping behavior in the context of tweeting about a different racial bias observed on another platform (Zoom). As might be expected, unintentional initiation is more likely to arise from the users' situated actions whereas more intentional initiation is more likely to arise from the actions of external auditors and not in the context of everyday use.

6.2 Awareness Raising

After detecting a problematic algorithmic behavior, the initiator or initiators broadcast what has been discovered to others. For example, in both Twitter cropping cases, this took the form of a tweet to other users. Others can then spread the word to even more people. Promoting the audit is a valuable part of the process, as it can bring more people into the discussion and increases visibility of the issue at hand. Sometimes, like in the racial cropping Twitter case, this promotion by other users creates the opportunity for further sharing and for participating in other parts of the audit like hypothesis forming and testing, which we explain in the next section. Other times, when discussion between users is more limited by a platform like in the Booking.com case where users had to communicate through reviews, this promotion by other users solely creates more awareness among those who see it, helping them to adjust their own behavior in using the system.

Raising awareness has ramifications both internal and external to the everyday algorithm audit. Internally, sharing information about the audit, for example in the racial cropping case on Twitter when some people created larger threads combining information from multiple places, serves to draw more participants into the audit in a cascading, self-referential way that potentially increases its collective nature. That is, the more users promote the audit, the more new people notice its happening and are then more likely to promote the audit themselves. Externally, sharing information about the audit makes it more likely that the awareness will eventually extend beyond the audit, as will be discussed more in Remediation below.

Everyday audit participants also can generate awareness through supporting the mechanisms of the audit. The LGBTQ+ YouTube demonetization is an example of this, as some YouTubers worked together to collect keywords, and a large number of people did a lot of YouTube tagging as well as shaming the company on Twitter for its actions. While this can promote the audit itself, these actions are more focused on assisting — but not guiding — other parts of the audit. Often a core group of people collaborate in the auditing process and are surrounded by a larger ring of people supporting their actions. The lack of many people in this support ring might help explain the lower traction for the Twitter gender cropping case as compared to the racial cropping case. Both the core group and the surrounding ring form a counterpublic space that supports the everyday audit as it progresses.

6.3 Hypothesizing & Testing

Following initiation and/or awareness raising, an everyday algorithm audit may progress toward users forming hypotheses for the observed behavior and intentionally testing an algorithmic system to collect further evidence. This process may involve just the individual who initiated the audit or may extend to many users if the initiator raised awareness and inspired others to get involved. Individual users who proceed alone continue with little support either from other situated users or from external experts. In more collective everyday algorithm audits, a group of everyday users organizes and builds on each others' efforts in the process. The users collectively decide on and steer the process, often through online discussion.

When users start hypothesizing about the potential biases of a system and testing their hypotheses, they usually develop "folk theories": non-authoritative theories a user forms to explain and understand how a technological system works [23, 27]. These theories, regardless of their validity, shape user behaviors and inform how they act around a system. In an everyday algorithm audit, these theories help users to form hypotheses about the sources of a bias, then test them. Sometimes these tests fail, such as switching the tie colors of McConnell and Obama in the Twitter racial cropping algorithm case to test the hypothesis that the algorithm might prioritize red over blue;[29]

---

[29]https://twitter.com/bascule/status/1307440596668182528

other times these tests succeed, such as manipulating the algorithm inputs in the Booking.com case to test the hypothesis that the algorithm skewed the minimum score up [28]. These examples show the importance of folk theory development in the process of an everyday audit: the more the platform allows users to develop and test theories, the more users have the opportunity to discover biases. One way to empower users in this process is *seamful* design; as Eslami et. al. described, adding visible hints disclosing aspects of an automation process, or seams, into the design of an algorithmic system can help users construct better conceptual understandings and theories about the system [27]. This increased transparency can provide users with more visibility, which may aid in detecting potential biases. The increasing number of everyday algorithm audits, along with new regulations around algorithmic bias [39, 49], can push algorithmic sociotechnical systems for a more transparent and seamful design.

Hypothesizing and testing often involves different degrees and forms of collaboration. In cases where users perform the auditing in an individual endeavor, they might spot an issue and quickly report it via social media channels or the platforms' internal discussion forum or, if coming from a background of high technical expertise, they might conduct the auditing in a more systematic way alone. In other more collective cases, users come together to make sense of how the system actually produces such biased results. During this process, they might raise new evidence, test each other's evidence, and collectively search for patterns that emerged from their discussion. For example, in Yelp's filtering algorithm case, users collectively worked together on the Yelp platform itself to form different hypotheses about how the algorithm works, test those hypotheses, and report the results back to the community.

Similar to the motivations behind initiation of the audit, the hypothesizing and testing process can range from organized to organic. More organized testing often involves external stakeholders (e.g., researchers, developers, journalists) who intervene in the process, while more organic testing often arises naturally without outside intervention as everyday users detect and assess algorithmic behaviors and play an active role in determining the course the audit takes. We argue that everyday algorithm auditing can receive — and often benefit from — different degrees and forms of external intervention. For instance, in some cases, developers with insight into the inner workings of a given system may be able to provide useful feedback to guide hypothesis formation, such as in the Twitter cropping case when an engineering team lead chimed in to let audit participants know that facial recognition is not part of the cropping algorithm.[30] To the extent that external interventions represent *guidance* for an otherwise collectively-led audit, rather than directing or orchestrating the collective's activities as in a crowdsourced/collective audit, we do not view such interventions as invalidating the *everyday* nature of the audit.

6.4 Remediation

At its core, an everyday algorithm audit has a singular objective: instigate change based on the issues identified and investigated. Change can occur at multiple levels. One type of change creates or increases external awareness through publicity. As the audit gains traction, news and other media sources are likely to become aware of the audit as well and may publish articles on the audit and the issues it highlights. This leads to broad social impact via greater awareness of the issues in the public sphere by all types of people, who may then choose to share this information with still others, alter their own patterns of behavior, or even take part in the audit if it is still occurring.

Another type of change comes in the form of legal action. For example, in the Yelp case, close to 700 lawsuits were filed in attempts to force Yelp to change its review filtering algorithm [52]. Though none of the lawsuits in the Yelp example were successful, they had the potential to put

---

[30]https://twitter.com/ZehanWang/status/1307461285811032066

the platform in a position in which it would have been required to respond and enact change or risk behaving illegally. There are also less direct repercussions of the legal action. Regardless of whether a lawsuit has the desired result for the plaintiff, the filing of a legal case in itself creates publicity for the issue.

Perhaps the most effective type of change is at the platform level. Sometimes when an everyday algorithm audit brings a problem to the forefront, the creators or employers of the algorithm take notice and agree that a change should occur. An official statement can create legitimacy around the audit and the identified issue, and swift redress can occur, since those who have decided on change now include those who have the power to enact change. Both of these resulted from the Twitter racial cropping case, as Twitter posted on the company blog [3] and committed to changing the way that image previews are cropped. In another example, Google revised its image recognition algorithm's racial bias in classifying some Black users as gorillas (which was found and reported by everyday users) [92]. However, platforms' attempts at redress are not always effective. For example, Google preemptively announced in 2020 that it had fixed its translation classifier to address bias that assigned gendered pronouns in an often stereotypical way to professions and activities when translating gender-neutral languages [66, 96]; the issue reemerged in 2021 with other languages.[31]

When platforms fail to act to fix a harmful and biased algorithmic behavior, everyday users sometimes go beyond detecting, testing, and reporting a bias to begin rectifying the behavior themselves. For example, in the Booking.com case, users started manipulating the algorithm to return lower algorithmically calculated ratings in their hotel reviews so that the actual ratings better aligned with users' intended ratings [28]. Everyday users also engage in what we call "collective repair" to resist and ameliorate the harm an algorithm can bring to a community. For example, in an everyday audit TikTok users found that the platform's For You Page algorithm suppresses content created by people with certain social identities, such as LGBTQ+ users and people of color. In response, some users worked together to amplify the suppressed videos by specifically engaging with them and adding comments [54, 82].

7 DESIGN IMPLICATIONS

We have discussed users' auditing behaviors around the Twitter image cropping algorithm and around rating algorithms on Yelp and Booking.com. Across these cases of everyday algorithm auditing, we observed similar patterns: a group of day-to-day users comes together to detect, understand, and interrogate problematic machine behaviors during the course of their ordinary interactions with algorithmic systems. In this section, we explore what lessons might be drawn from the study of everyday algorithm auditing to overcome limitations of previously proposed auditing approaches. We ask how designers might facilitate everyday algorithm auditing practices, building upon regular users' existing motivations to engage in such auditing behaviors and supporting their unique strengths in surfacing harmful algorithmic behaviors. How might designers support and scaffold everyday algorithm audits, without intervening so heavily as to sacrifice their bottom-up, user-driven nature?

Below we outline five broad categories of potential design interventions — (a) community guidance, (b) expert guidance, (c) algorithmic guidance, (d) organizational guidance, and (e) incentivization — and discuss potential design trade-offs.

---

[31]e.g., https://twitter.com/frostaf/status/1376699332917870592, https://twitter.com/imajeanpeace/status/1374135940898156544, & https://twitter.com/DoraVargha/status/1373211762108076034

7.1 Design for Community Guidance

Across our cases, we observed different degrees and forms of collaboration among everyday auditors. In general, more collective efforts may lead to more impactful everyday algorithm audits. For example, everyday auditors in the Yelp case engaged in more collaborative auditing behaviors compared with the Booking.com case. We could speculate that, through increased interaction with other everyday auditors, Yelp users may have been more likely to form "counterpublics" against the dominant algorithmic system, in a collective effort toward sensemaking and consensus building. Moreover, more collective everyday auditing efforts may be more successful in raising awareness and pushing for changes on the algorithmic, platform, and even societal level, as we observed in our discussions around Twitter's cropping algorithm.

One way designers might support collective auditing behavior is to offer mechanisms that help everyday auditors guide each other's efforts. At minimum, this might include providing spaces for community discussion so that everyday auditors can discuss, raise awareness about, and build upon each other's findings, as we have already observed in the Yelp and Twitter cases. Apart from directly facilitating the auditing process, public discussion forums may help garner support from advocacy groups and news organizations to raise awareness and push for potential remediations on different levels, as we observed in the YouTube LGBTQ+ demonetization case.

Beyond simply providing spaces for discussion, we might imagine tailoring or augmenting discussion channels with mechanisms that are explicitly designed to support everyday auditing. For example, a discussion platform designed for everyday auditing might allow community members to upvote specific algorithmic behaviors that other community members have reported, in order to collectively surface the most *severe* reports, or reports that may *benefit from further discussion*. Furthermore, when an everyday auditor selects a particular report, the discussion platform might provide an overview of relevant hypotheses and findings that other auditors have generated so far, to help the auditor build upon previous contributions instead of retreading old ground.

When designing to support community guidance, we should also consider the possibilities of leveraging existing platforms rather than creating entirely new ones. In the Yelp case, users were able to leverage embedded discussion forums to organize their everyday auditing. In the Twitter case, the platform being audited was itself a discussion platform, allowing users to discuss and test algorithmic behavior *simultaneously*. This evokes Grevet and Gilbert's "piggyback prototyping" technique in which existing, popular platforms (e.g., Facebook and Twitter) are appropriated in the design of new social computing systems in an attempt to help overcome the challenges of obtaining critical mass on entirely new platforms [40]. A potential trade-off is that hosting community discussion *within* the same platforms that are under audit may lead to concerns about platform-driven censorship or suppression, given uneven power dynamics and the potential for value conflicts between justice-oriented audits and profit-oriented platforms. We anticipate that in certain cases, everyday auditors may feel more comfortable hosting these discussions elsewhere.

7.2 Design for Expert Guidance

Although the cases we reviewed were often initiated organically by situated users, others with relevant technical expertise sometimes weighed in as well. In the Twitter photo cropping case, developers from Twitter intervened in the discussion of racial bias to weigh in with some of the background knowledge they had available on how the cropping algorithm functions.[32] We expect that everyday algorithm audits can benefit from expert input, whether from system developers, researchers, or regulators. However, it remains an open question what forms and degrees of expert engagement may be most helpful in facilitating everyday audits.

---

[32] https://twitter.com/ZehanWang/status/1307461285811032066

To help steer everyday auditors' ongoing efforts in more productive directions, designers might invite relevant experts to continuously monitor community discussions and intervene as needed [13, 26]. For example, at any stage of an everyday audit, a product team might weigh in and provide feedback, based on technical knowledge of their products, regarding the plausibility of user-generated hypotheses. However, if this is not handled carefully, experts' involvement may risk diminishing everyday auditors' sense of community and autonomy. For example, the insertion of technical or domain experts into the middle of an everyday audit might create unequal power dynamics in community discussions.

Relevant experts may also provide guidance to users at the initiation stage. For example, to complement their own internal auditing efforts, a product team may actively draw users' attention to suspected issues or components of a system that they would like users to audit, then suggest testing strategies for users to try [48]. Similarly, in the case of ImageNet Roulette [19], researchers built an interface to help everyday users quickly and easily audit a model trained on the ImageNet dataset, then disseminated this interface on social media. In turn, the ImageNet Roulette interface went viral, spurring a large and diverse group of users to test for algorithmic biases using images of their choosing. Although audits were initiated by technical experts in both of the above examples, users still had agency and autonomy in directing their collective efforts via social media.

The design of bidirectional feedback mechanisms between developers and everyday users of algorithmic systems represents a critical direction for future research. In addition to the need for systems that help developers provide effective feedback and guidance to users who are engaged in everyday auditing behaviors, there are opportunities to design systems that scaffold users in providing more useful, actionable feedback to developers.

7.3 Design for Algorithmic Guidance

For certain auditing tasks, such as exploring a system's behavior, gathering evidence, or testing hypotheses, everyday auditors may benefit from algorithmic assistance. For instance, designers might develop auditing interfaces that automatically surface potentially important instances for everyday auditors to examine further. As a foundation for such interfaces, it may be possible to build upon emerging algorithmic techniques for crowd-in-the-loop detection of "unknown unknowns" in ML models (e.g., [6, 56, 59, 64, 87, 101]), which automatically surface cases that are more likely to be mislabelled and/or misclassified. These methods focus on surfacing regions of a model's error space in which the model is highly confident yet incorrect [56]. This form of algorithmic guidance could supplement guidance that users may receive from each other (e.g., via upvoting mechanisms) and from expert guidance (e.g., suggestions for how everyday auditors should direct their search efforts), potentially surfacing cases that these other forms of guidance would miss.

In some of the everyday auditing cases we have reviewed in this paper, establishing the presence of certain issues requires some form of *comparison*. We distinguish between two major categories of auditing behaviors: "instance-based" and "comparison-based" auditing. Instance-based auditing applies to cases where the observation of an isolated harmful algorithmic behavior is sufficient to ground an argument for remediation. For example, when users found that the Google Photos app assigned a photo of Black people with the label "gorillas" [42], it was not necessary to demonstrate that most photos of Black people receive the same label, nor was it necessary to demonstrate that Black people are consistently assigned this label at a higher rate than white people. Rather, users felt that the app should *never* assign such a harmful label to images of Black people. By contrast, comparison-based auditing applies to cases where, in order to establish that a harmful bias truly exists and is worth addressing, it is necessary to compare multiple instances and demonstrate the existence of statistical patterns. For example, when Sweeney [84] found that a web search for her own name yielded arrest record ads, she wanted to understand whether systematic discrimination

was present. To do so, it was necessary to search for many Black sounding names and statistically compare the results against those for other names [84]. To support such forms of comparative auditing, designers might provide interfaces and algorithmic guidance to empower everyday auditors, who typically lack knowledge of statistics or causal modeling, in testing statistical or causal hypotheses.

7.4 Design for Organizational Guidance

In our cases, we also observed fluid divisions of labor among users across the lifetime of an everyday algorithm audit. For example, in the Twitter cropping case, some users largely focused on tweeting images to test how the image cropping algorithm would behave for each. Meanwhile, other users contributed more in retweeting and raising awareness about observed issues, or in forming hypotheses about observed algorithmic behaviors.

Designers could help guide the division of labor through interventions that encourage everyday auditors to take on particular roles during an auditing process, including roles that are expected to be important and might not arise organically. For example, across all of our cases, a major challenge was a lack of effective synthesis mechanisms to help users keep track of each other's findings and hypotheses and build upon each other's contributions so far (cf. [18]). "Synthesizer" may be a valuable role for some everyday auditors to take on, in order to help the collective make more cumulative progress. More broadly, we could imagine organizing everyday auditors along a number of major roles, such as (1) *Initiators*, who focusing on identifying problematic algorithmic behaviors; (2) *Generalizers*, who examinine instances flagged by initiators, and use these to inform hypotheses about the scope of the issue and possible underlying causes; (3) *Amplifiers*, who help share and broadcast what has been discovered to others, raising awareness of findings, hypotheses, and relevant discussions so far; and (4) *Synthesizers*, who transform the outputs of this iterative auditing process into summaries for others to build on.

7.5 Design for Incentivization

Finally, while community, expert, algorithmic, and organizational guidance may help to improve ongoing auditing processes, another major design opportunity is to explore mechanisms that incentivize everyday auditors to have more active and sustained participation. For example, one potential design direction is to offer an "algorithmic bias bounty program," drawing inspiration from "bug bounty" programs in cybersecurity. The core idea of security's "bug bounty" program is to incentivize hackers to share vulnerabilities with legitimate organizations for monetary and reputational rewards, as an alternative to selling or exploiting those vulnerabilities [67, 76]. Similarly, in the context of everyday algorithm audits, we could encourage users to directly report detected biases to organizations and platforms, offering social incentives (e.g., reputations and points) and monetary incentives accordingly. A potential risk is that, if not designed carefully, the provision of extrinsic rewards for algorithm auditing may reduce useful spontaneity or may inadvertently diminish users' existing intrinsic motivations to engage in everyday auditing behaviors.

Indeed, while this paper was under review, Twitter announced its own "bias bounty" program, inspired by cybersecurity "bug bounty" programs and by its own image cropping cases. This program aimed at incentivizing people to detect, test, and report potential biases [16]. However, such "bias bounty" programs also heavily rely on participants with relevant technical expertise to detect biases, similar to bug bounty programs in cybersecurity, which rely on experienced hackers to identify bugs. We see design opportunities here to develop systems that can combine strengths of both technical experts and everyday users. For example, an audit could begin by inviting a wide and open range of everyday users to participate in surfacing harmful algorithmic behaviors. Initial hypotheses and observations from everyday users could then serve to inform further, systematic

investigations by machine learning experts and analysts. The initial, open phase of auditing may uncover issues that a narrower group of technical experts might otherwise have missed. Indeed, in some ways, this is how the Twitter "bias bounty" program proceeded, given that participants in the program were directly informed by observations and hypotheses that everyday users had developed previously.

In addition to incentivizing auditors, there is also a need to incentivize developers and platforms both to implement these measures (e.g., supporting user-led audits on embedded discussion forums) and to remediate issues that users uncover. Holstein et al. (2019) found that commercial product teams, fearful of negative public attention, are often motivated to uncover harmful algorithmic behaviors as early in the process as possible. Acknowledging the limitations of existing auditing approaches, several industry practitioners interviewed in this research expressed interest in implementing mechanisms for user-led auditing [48]. Some product teams were motivated enough to experiment with developing their own collective auditing tools and processes. While these findings highlight that some product teams and companies are already motivated to implement such measures, others may be hesitant to encourage users to scrutinize their systems. Prior work has demonstrated that, even where motivation is lacking, public awareness raising around harmful behaviors in commercial AI systems can motivate target companies to prioritize remediation [70]. Similarly, recent work from Vincent et al. [93] discusses potential ways for the public to withhold their data contributions as leverage, reducing the effectiveness of specific data-driven technologies in order to motivate companies to address users' concerns. Finally, given the uneven power dynamics between users and platforms, structural interventions are also needed (e.g., policy and legal interventions) to ensure external accountability.

8 DISCUSSION

We have taken initial steps to theorize and explore an under-studied phenomenon in which everyday users detect, interpret, question, and bring attention to problematic machine behaviors in their day-to-day interactions with algorithms. We argue that such "everyday algorithm auditing" is especially powerful in detecting harmful behaviors that are challenging for existing auditing approaches discussed in the academic literature. In this section, we discuss the unique power and contributions of everyday algorithm auditing and raise a few open questions on how to better support it for future research.

8.1 The Power of Everyday Algorithm Auditing

Past research has highlighted the limitations and blindspots of existing auditing approaches and speculated that everyday users might be able to overcome these limitations (e.g., [29, 48]). In this paper, we take a step further by studying how such user-led audits unfold in the real world and by exploring how we might better support these processes. Through detailed case studies, we have shown that everyday users are able to detect harmful algorithmic behaviors that are challenging for traditional auditing approaches and that users can exercise agency and autonomy in directing their own auditing process. Below we highlight a number of unique contributions of everyday algorithm auditing and how it can help overcome some of the limitations presented in existing approaches.

First, everyday algorithm auditing leverages the *lived experiences* of everyday users. Past literature suggests that existing approaches often fail to involve auditors with relevant cultural backgrounds and lived experience that are critical to detect sensitive harms [99]. Everyday algorithm auditing, with contextually situated users, is especially powerful to appropriately identify the problems at hand. Although a crowdsourced/collaborative auditing approach [74] also invites users' participation, it often relies on crowdworkers, who do not necessarily have lived experiences with the

algorithms being audited, do not communicate and work together toward a common goal, and therefore might not be sensitive to the related harms.

Second, everyday algorithm auditing harnesses the *situated knowledge* of everyday users. As past literature suggests, many harmful machine behaviors are challenging to detect outside of situated contexts of use [17, 33, 48, 60, 78]. Through their day-to-day interactions with an algorithmic system, everyday users are particularly well positioned to detect these types of behaviors that emerge in real-world contexts of use, in the presence of complex social dynamics, and in the changing norms and practices of using algorithmic systems over time.

Third, in an everyday algorithm audit, users are able to form *counterpublics* via different social media channels or community forums and participate in their own collective sensemaking and consensus building. This is especially powerful since only in such collective attempts, will users be able to build on each others' contributions, test each others' hypotheses, and support each other in different forms (e.g., helping publicize their efforts). In contrast, in a crowdsourced/collaborative audit [74], users are often working on individualized tasks that have been assigned to them, without participating in discussion.

Finally, in an everyday algorithm audit, users have *control* over the auditing process. This differs from a crowdsourced/collaborative audit [74], in which users are hired via crowdsourcing platforms to work on decomposed subtasks determined entirely by outside parties; in an everyday algorithm auditing, users — often collectively — decide their own course of action. Their autonomy and agency help steer the audit in the directions that are most useful and meaningful for them, which are often overlooked by existing approaches.

8.2 Appropriate and Timely Interventions

We have discussed five types of design interventions we might offer to support everyday algorithm auditing: community guidance, algorithmic guidance, expert guidance, organizational guidance, and incentivization. We have also discussed some potential trade-offs that we need to consider when offering these types of interventions. Below we raise open questions regarding the appropriate timing and proper degree of intervention for future research.

8.2.1 When to intervene and when to stop. We have discussed that interventions can happen at any stage of an everyday algorithm audit and that different stages might benefit from different types of interventions. For example, we might anticipate that having expert guidance during the initiation stage (e.g., as in the case of ImageNet Roulette) and the hypothesizing and testing stage (e.g., in the Twitter racial cropping case) might be especially helpful. On the other hand, community guidance might be more useful at the awareness raising stage (e.g., as in the YouTube LGBTQ+ demonetization case) and the hypothesizing and testing stage (e.g., in the Yelp and Twitter cases). However, currently we have very limited understanding of the appropriate timing for intervention. For example, how can we maintain the organic nature of everyday algorithm auditing, especially at the initiation stage, but also prompt a group of people to start auditing first, which might inevitably remove some of the organic nature?

8.2.2 How much intervention. We have also discussed different forms of intervention and some hypothesized trade-offs. For example, introducing expert guidance might create unequal power dynamics in community discussions and risk diminishing users' sense of community and autonomy. Hosting community discussions within the same platforms that are under audit may lead to concerns about platform-driven censorship or suppression. Currently we are only at the starting point to understand and explore the appropriate degrees of these interventions. How much intervention is too much? What are the trade-offs in moving from more organic to more organized at

different phases of an everyday audit? These questions need further research and investigation; the exploration of everyday algorithm auditing is in its infancy, and our work presents a first step .

9 CONCLUSIONS

In this paper, we have drawn on real-world case studies and prior theories of everyday resistance and counterpublics to understand how everyday users — either individually or collectively — detect, understand, and interrogate problematic machine behaviors in their day-to-day interactions with algorithmic systems. By comparing the lifetime and dynamics of several cases of everyday algorithm auditing, we have proposed a process-oriented view of everyday algorithm audits, involving initiation, awareness raising, testing and hypothesizing, and remediation. We have drawn lessons from these cases for future research and design around everyday algorithm auditing, outlining five broad categories of potential design interventions to support everyday audits — (a) community guidance, (b) expert guidance, (c) algorithmic guidance, (d) organizational guidance, and (e) incentivization — along with potential trade-offs.

As the first work conceptualizing and studying everyday algorithm auditing, this research is highly exploratory in nature. The lifetime and dynamics presented as part of an everyday audit are intended to be representations of what we see at this stage of exploration, capturing both current practices we have observed and possibilities we see for future practice. As such, these represent a set of untested assumptions about everyday algorithm audits that could change over time as more empirical work is conducted in this space. That is, neither the cases presented in this paper nor the approaches we used to analyze them are comprehensive: they outline an observed phenomenon and serve as illustrative examples to better understand how everyday audits work, what paths everyday audits could take, and what impacts everyday audits could have. Future research should conduct in-depth follow-up studies with users to investigate the practices and strategies they use to find and make sense of harmful algorithmic behaviors. Another fruitful, complementary area for future research includes an exploration of the design opportunities discussed above. For example, how can we design platforms that support users in conducting more effective everyday algorithm audits, both individually and collectively?

In sum, we view this work as an initial step toward bridging the gap between algorithm auditing approaches in academia/industry and everyday auditing behaviors that emerge in day-to-day use of algorithmic systems. It is our hope that this work will help to inform future research on everyday algorithm auditing, as well as the design of future platforms and tools that empower people to meaningfully audit the algorithmic systems that impact their lives every day.

REFERENCES

[1] 2020. Twitter investigates racial bias in image previews. *BBC News* (Sep 2020). https://www.bbc.com/news/technology-54234822

[2] 2021. The Algorithmic Justice League: Mission, Team and Story. https://www.ajl.org/about

[3] Parag Agrawal and Dantley Davis. 2020. Transparency around image cropping and changes to come. *Twitter* (2020). https://blog.twitter.com/en_us/topics/product/2020/transparency-image-cropping.html.

[4] Michael Anderson and Jeremy Magruder. 2012. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal* 122, 563 (2012), 957–989.

[5] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing race and gender discrimination in online housing markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 24–35.

[6] Gagan Bansal and Daniel Weld. 2018. A coverage-based utility model for identifying unknown unknowns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[7] Todd Bishop. 2015. Amazon changes its key formula for calculating product ratings and displaying reviews. *GeekWire* (June 2015). https://www.geekwire.com/2015/amazon-changes-its-influential-formula-for-calculating-product-ratings

[8] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050* (2020).

[9] Rucksack Brian. 2019. Major changes to review scoring method on Booking com. https://hostelmanagement.com/forums/major-changes-review-scoring-method-booking-com.html

[10] Khristopher J Brooks. 2020. Twitter users say the platform crops out Black faces. *CBS News* (Sep 2020). https://www.cbsnews.com/news/twitter-image-cropping-algorithm-racial-profiling/

[11] Joy Buolamwini. 2016. How I'm fighting bias in algorithms. *November, TEDx Beacon Street)[Video File]* (2016).

[12] 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.

[13] Joel Chan, Steven Dang, and Steven P Dow. 2016. Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1223–1235.

[14] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[15] Zhilong Chen, Hancheng Cao, Fengli Xu, Mengjie Cheng, Tao Wang, and Yong Li. 2020. Understanding the Role of Intermediaries in Online Social E-commerce: An Exploratory Study of Beidian. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.

[16] Rumman Chowdhury and Jutta Williams. 2021. Introducing Twitter's first algorithmic bias bounty challenge. *Twitter* (2021). https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge

[17] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and addressing algorithmic bias in practice. *Interactions* 25, 6 (2018), 58–63.

[18] Justin Cranshaw and Aniket Kittur. 2011. The polymath project: lessons from a successful online collaboration in mathematics. In *Proceedings of the 2011 CHI conference on Human Factors in Computing Systems*. 1865–1874.

[19] Kate Crawford and Trevor Paglen. 2021. Excavating AI: The politics of images in machine learning training sets. *AI & Society* (2021), 1–12.

[20] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (Oct. 2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[21] Richard Dawkins. 2016. *The selfish gene.* Oxford University Press.

[22] Michel De Certeau. 1984. *The practice of everyday life, trans Steven Rendall.* Berkeley: University of California Press.

[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*. 1–12.

[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*. 3163–3174.

[25] Nicholas Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. *The Tow Center for Digital Journalism* (2014).

[26] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. 1013–1022.

[27] 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*. 2371–2382.

[28] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; things can be worse than they appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.

[29] 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*. 1–14.

[30] Kristie Fisher, Scott Counts, and Aniket Kittur. 2012. Distributed sensemaking: Improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems*. 247–256.

[31] Geoffrey A. Fowler. 2011. Judge dismisses suit against Yelp. *The Wall Street Journal* (Oct. 2011). https://www.wsj.com/articles/SB10001424052970204505304577002170423750412

[32] Nancy Fraser. 1990. Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social text* 25/26 (1990), 56–80.

[33] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.

[34] Brian Fung. 2015. The Internet has unearthed more racist Google Maps results. *The Washington Post* (May 2015). https://www.washingtonpost.com/news/the-switch/wp/2015/05/21/the-internet-has-unearthed-more-racist-google-maps-results

[35] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.

[36] R Stuart Geiger and Aaron Halfaker. 2013. Using edit sessions to measure participation in Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. 861–870.

[37] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167, 2014 (2014), 167.

[38] Eric Goldman. 2014. Court says Yelp doesn't extort businesses. *Forbes* (Sept. 2014). https://www.forbes.com/sites/ericgoldman/2014/09/03/court-says-yelp-doesnt-extort-businesses

[39] Bryce Goodman and Seth Flaxman. 2016. EU regulations on algorithmic decision-making and a "right to explanation". In *ICML Workshop on Human Interpretability in Machine Learning*.

[40] Catherine Grevet and Eric Gilbert. 2015. Piggyback prototyping: Using existing, large-scale social computing systems to prototype new ones. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*. 4047–4056.

[41] Rosanna E. Guadagno, Daniel M. Rempala, Shannon Murphy, and Bradley M. Okdie. 2013. What makes a video go viral? An analysis of emotional contagion and Internet memes. *Computers in Human Behavior* 29, 6 (2013), 2312–2319. https://doi.org/10.1016/j.chb.2013.04.016

[42] Jessica Guynn. 2015. Google photos labeled black people 'gorillas'. *USA Today* (June 2015). https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465

[43] Jürgen Habermas. 1989. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press.

[44] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. 163–172.

[45] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*. 305–318.

[46] Alex Hern. 2020. Twitter apologises for 'racist' image-cropping algorithm. *The Guardian* (Sept. 2020). https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm

[47] Kenneth Holstein, Erik Harpstead, Rebecca Gulotta, and Jodi Forlizzi. 2020. Replay enactments: Exploring possible futures through historical data. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1607–1618.

[48] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[49] White House. 2016. Big Data: A Report on Algorithmic Systems. *Opportunity, and Civil Rights* (2016).

[50] Khari Johnson. 2019. AI predictions for 2019 from Yann LeCun, Hilary Mason, Andrew Ng, and Rumman Chowdhury. *VentureBeat* (Jan 2019). https://venturebeat.com/2019/01/02/ai-predictions-for-2019-from-yann-lecun-hilary-mason-andrew-ng-and-rumman-chowdhury/

[51] Khari Johnson. 2020. Apparent racial bias found in Twitter photo algorithm. *VentureBeat* (Sep 2020). https://venturebeat.com/2020/09/20/apparent-racial-bias-found-in-twitter-photo-algorithm/

[52] Inkoo Kang. 2013. Businesses: 'Yelp is the thug of the Internet'. *MuckRock* (2013). https://www.muckrock.com/news/archives/2013/jan/23/businesses-yelp-thug-of-the-internet.

[53] Inkoo Kang. 2013. Read nearly 700 FTC complaints regarding Yelp. *MuckRock* (Jan 2013). https://www.muckrock.com/news/archives/2013/jan/23/businesses-yelp-thug-of-the-internet/

[54] Nadia Karizat, Daniel Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic folk theories and identity: How TikTok users co-produce Knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction* CSCW2 (2021), forthcoming.

[55] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 417–432.

[56] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[57] Michelle Seng Ah Lee and Jatinder Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[58] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 497–506. https://doi.org/10.1145/1557019.1557077

[59] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. 2020. Towards hybrid human-AI workflows for unknown unknown detection. In *Proceedings of The Web Conference 2020*. 2432–2442.

[60] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[61] Ivan Mehta. 2021. Why Twitter's image cropping algorithm appears to have white bias. *TNW | Neural* (Mar 2021). https://thenextweb.com/news/why-twitters-image-cropping-algorithm-appears-to-have-white-bias

[62] Cade Metz. 2019. 'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You? *The New York Times* (Sep 2019). https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html

[63] Ronald B Mincy. 1993. The Urban Institute audit studies: Their research and policy context. In *Clear and convincing evidence: Measurement of discrimination in America*, M. Fix and R. J. Struyk (Eds.). Urban Institute Press, 165–86.

[64] Mengjun Ni, Jing Yang, Xin Lin, and Liang He. 2017. Reducing unknown unknowns with guidance in image caption. In *International Conference on Artificial Neural Networks*. Springer, 547–555.

[65] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism.* NYU Press.

[66] Parmy Olson. 2018. The algorithm that helped google translate become sexist. *Forbes* (Feb. 2018). https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/?sh=6c1d0807daa2

[67] Andy Ozment. 2004. Bug auctions: Vulnerability markets reconsidered. In *Third Workshop on the Economics of Information Security*. 19–26.

[68] Francesca Palmiotto. 2020. Challenging automated filtering systems - The case of Yelp. https://ai-laws.org/en/2020/05/challenging-automated-filtering-systems-the-case-of-yelp/

[69] Charles C. Ragin and Howard S. Becker. 1992. *What is a case?: Exploring the foundations of social inquiry.* Cambridge University Press.

[70] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

[71] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.

[72] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[73] Aja Romano. 2019. A group of YouTubers is trying to prove the site systematically demonetizes Queer content. *Vox* (Oct. 2019). https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report

[74] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* (2014).

[75] Satviki Sanjay. 2020. Twitter's AI Revealed to Have Racial Bias After a Viral Photo Experiment. *Vice* (Sep 2020). https://www.vice.com/en/article/93547v/twitter-ai-algorithm-has-racial-bias

[76] Stuart Schechter. 2002. How to buy better testing using competition to get the most security and robustness for your dollar. In *International Conference on Infrastructure Security*. Springer, 73–87.

[77] James C Scott. 1985. *Weapons of the weak: Everyday forms of peasant resistance.* Yale University Press.

[78] Nick Seaver. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4, 2 (2017). https://doi.org/doi:10.1177/2053951717738104

[79] Nick Seaver. 2019. Knowing algorithms. *Digital STS* (2019), 412–422.

[80] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 59–68.

[81] Limor Shifman. 2013. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication* 18, 3 (2013), 362–377.

[82] Ellen Simpson and Bryan Semaan. 2021. For you, or For "you"? Everyday LGBTQ+ encounters with TikTok. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–34.

[83] Lucy A Suchman. 1987. *Plans and situated actions: The problem of human-machine communication.* Cambridge University Press.

[84] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10–29.

[85] Lucas Theis and Zehan Wang. 2018. Speedy Neural Networks for Smart Auto-Cropping of Images. https://blog.twitter.com/engineering/en_us/topics/infrastructure/2018/Smart-Auto-Cropping-of-Images.html

[86] Harry Tucker. 2016. Australian Uber drivers say the company is manipulating their ratings to boost its fees. *Businessinsider* (May 2016). https://www.businessinsider.com.au/australian-uber-drivers-say-the-company-is-manipulating-their-ratings-to-boost-the-companys-fees-2016-5

[87] Colin Vandenhof and Edith Law. 2019. Contradict the Machine: A Hybrid Approach to Identifying Unknown Unknowns. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2238–2240.

[88] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[89] Julia Velkova and Anne Kaun. 2019. Algorithmic resistance: Media practices and the politics of repair. *Information, Communication & Society* (2019), 1–18.

[90] Neil Vigdor. 2019. Apple card investigated after gender discrimination complaints. *The New York Times* (Nov. 2019). https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

[91] James Vincent. 2016. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge* (Mar 2016). https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

[92] James Vincent. 2018. Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. *The Verge* (Jan. 2018). https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

[93] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. 2021. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 215–227.

[94] Shuaishuai Wang. 2020. Calculating dating goals: data gaming and algorithmic sociality on Blued, a Chinese gay dating app. *Information, Communication & Society* 23, 2 (2020), 181–197.

[95] Karl E Weick. 1995. *Sensemaking in organizations*. Vol. 3. Sage.

[96] Kyle Wiggers. 2020. Google debuts AI in Google Translate that addresses gender bias. *VentureBeat* (Apr 2020). https://venturebeat.com/2020/04/22/google-debuts-ai-in-google-translate-that-addresses-gender-bias

[97] Sijia Xiao, Danaë Metaxa, Joon Sung Park, Karrie Karahalios, and Niloufar Salehi. 2020. Random, messy, funny, raw: Finstas as intimate reconfigurations of social media. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[98] Yelp. 2010. Yelp's Review Filter Explained. https://blog.yelp.com/2010/03/yelp-review-filter-explained

[99] Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward inclusive tech policy design: A method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology* 21, 2 (2019), 89–103.

[100] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist—it's time to make it fair. *Nature* 559 (2018), 324–326.

[101] Ángel Alexander Cabrera, Abraham Druck, Jason I Hong, and Adam Perer. 2021. Discovering and Validating AI Errors With Crowdsourced Failure Reports. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* 5 (2021), forthcoming. https://doi.org/10.1145/3479569