How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services

MORGAN KLAUS SCHEUERMAN, JACOB M. PAUL, and JED R. BRUBAKER, University of Colorado Boulder

Investigations of facial analysis (FA) technologies—such as facial detection and facial recognition—have been central to discussions about Artificial Intelligence's (AI) impact on human beings. Research on automatic gender recognition, the classification of gender by FA technologies, has raised potential concerns around issues of racial and gender bias. In this study, we augment past work with empirical data by conducting a systematic analysis of how gender classification and gender labeling in computer vision services operate when faced with gender diversity. We sought to understand how gender is concretely conceptualized and encoded into commercial facial analysis and image labeling technologies available today. We then conducted a two-phrase study: (1) a system analysis of ten commercial FA and image labeling services and (2) an evaluation of five services using a custom dataset of diverse genders using self-labeled Instagram images. Our analysis highlights how gender is codified into both classifiers and data standards. We found that FA services performed consistently worse on transgender individuals and were universally unable to classify non-binary genders. In contrast, image labeling often presented multiple gendered concepts. We also found that user perceptions about gender performance and identity contradict the way gender performance is encoded into the computer vision infrastructure. We discuss our findings from three perspectives of gender identity (self-identity, gender performativity, and demographic identity) and how these perspectives interact across three layers: the classification infrastructure, the third-party applications that make use of that infrastructure, and the individuals who interact with that software. We employ Bowker and Star's concepts of "torque" and "residuality" to further discuss the social implications of gender classification. We conclude by outlining opportunities for creating more inclusive classification infrastructures and datasets, as well as with implications for policy.

CCS Concepts: • **Social and professional topics → User characteristics**; *Gender*.

Additional Key Words and Phrases: categorization, computer vision, facial analysis, image labeling, facial recognition, facial detection, Instagram, identity, gender.

ACM Reference Format:
Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 144 (November 2019), 33 pages. https://doi.org/10.1145/3359246

1 INTRODUCTION

Artificial Intelligence (AI) has quickly become intertwined with both the most mundane and the most critical aspects of human life. Computerized systems handle everything from personalized

Authors' address: Morgan Klaus Scheuerman, morgan.scheuerman@colorado.edu; Jacob M. Paul, jacob.paul@colorado.edu; Jed R. Brubaker, Jed.Brubaker@colorado.edu, University of Colorado Boulder, Boulder, Colorado, 80309.

2573-0142/2019/11-ART144 $15.00
https://doi.org/10.1145/3359246

recommendations [20] to dictating vacation rental pricing [64] to determining a previously convicted individual's likelihood to relapse into criminal behavior [32]. Algorithmic decision-making often travels through a layered and interconnected pipeline of algorithmic infrastructure. These pipelines are often weaved together into a network of inputs and outputs, databases and data, all for the purposes of solving tasks efficiently.

As AI increasingly intersects with human life, characteristics of *human identity* are operationalized and made computable. A spectrum of human characteristics—such as age, race, and gender—are being discretely separated and embedded into algorithmic computer systems in new and unique ways: without "user" input. Computers are being trained to recognize and respond to human characteristics. The use cases abound: collecting demographic [27, 116] and user profile information [22] to recommend personalized content; using gender, age, and blood pressure to predict the risk of heart disease [104]; and even automatically detecting signs of mental illness in user social media images [55, 91, 109].

Image detection and classification, in particular, represents a pertinent domain where we see a tight coupling of human identity and computation. Perhaps the most salient example is *automated facial analysis technology* (FA) [19], an umbrella term for computer vision methods that use machine learning (ML) techniques to automate problems related to reading the human face [63]. FA is often discussed in the context of two specific tasks: facial detection and facial recognition. Both use computational methods to measure the human face, whether simply to detect that a face is present (i.e., facial detection) or to detect a specific individual's face (i.e., facial recognition).

Identity classification in automated FA has become a site of social computing research in recent years. Specifically, researchers have started to examine *automatic gender recognition (AGR)*, an FA task for classifying the gender of human subjects (for overviews see [46, 57]). Evaluation studies are increasingly common. For example, Buolamwini and Gebru evaluated FA services from Microsoft, IBM, and Face++ and found higher gender classification error rates for dark-skinned women than for white men [19]. Muthukumar et al. sought to isolate the reason that facial analysis systems often worked disproportionately worse on women of color than other groups, highlighting lip, eye, and cheek structure as a primary predictor for gender in FA systems [76]. Within CSCW and HCI, one major thread of scholarship addresses concerns over how gender is conceptualized. Existing work has discussed the potential harms when gender is collapsed into a single gender binary—"male" or "female"—rather than approaching gender as socially constructed, non-binary, or even fluid. Hamidi et al. conducted a qualitative examination of transgender and/or non-binary[1](which we abbreviate to "trans" from here on) [30] users' expectations and impressions of AGR systems, uncovering widespread concern about the ramifications of AGR [46]. Similarly, Keyes conducted a systematic literature review of academic AGR research, highlighting the potentially negative implications that binary algorithmic classifications could have on trans individuals through the reinforcement of systematic marginalization [57].

While this previous work is invaluable to the design of human-centered algorithms moving forward, the issues raised by FA and AGR are present now and pressing. Popular press is replete with accounts of potential and actual misuse of facial analysis. These include cautions about the risk of providing facial analysis to law enforcement [52], and the potential for abuse when such services are adopted by airport or immigration systems [72]. As computer vision has been added to the cloud-based infrastructure services used by third-party developers, questions about how to manage responsibility and oversight have quickly followed. Microsoft, for example, has argued for

[1]We use the term "trans and/or non-binary" to acknowledge and respect both non-binary individuals who *do* identify as trans and non-binary individuals who *do not* identify as trans. We acknowledge this difference in our shortened umbrella use of "trans" as well.

more regulation that will clarify whether the service providers or third-party applications should bare ultimate responsibility for any misuse [17], though their stance has been widely criticized [105].

In our research, we speak to these concerns by studying how computers *see* gender. We intentionally focused our attention on cloud-based providers of computer vision services, focusing on how these services operationalize gender and presented as a feature to third-party developers. Specifically, we focused on answering the following research questions:

(1) How do commercial FA services codify gender characteristics (through facial classification and labels)?
(2) How accurate are commercial FA services at classifying images of diverse genders (including binary and non-binary genders)?
(3) How do individuals self-describe internal gender identity and how does this compare with the descriptions FA infrastrutures provide?

To answer these research questions, we present results from a two-phase study. We start by surveying related work to situate our contribution. We focus on three areas of scholarship in social computing that contribute to understanding the intersection of gender identity and facial analysis: identity, infrastructure, and facial analysis technology. We then share results from a technical analysis of commercial facial analysis and image labeling services, focusing on how gender is embedded and operationalized in these services. Building on our technical analysis, we then present our evaluation study of five services. Using a manually constructed image dataset of 2450 faces with diverse genders from Instagram, we conducted a performance evaluation to determine the success rate of commercial classifiers across multiple genders. We then share our analysis of image labeling services, focused on how gender is detected by labeling services and embedded in the labels they provide. Finally, we compared these services with the content Instagram users provided in their own captions and hashtags, revealing clashes between social and technical perspectives about gender identity.

We reflect on our findings in relationship to three different perspectives of gender: internal self-held gender, gender performativity, and systematic demographic gender. We discuss how these three different perspectives emerge and are omitted through layers of infrastructure and third-party applications, resulting in people experiencing what Bowker and Star call *torque*, especially when they reside in the *residual* spaces that are unrecognizable to these systems.

If researchers are going to propose design and policy recommendations, it is critical to understand how gender is currently classified in available commercial services. Our findings build on previous scholarship to provide an empirical analysis of FA services. Where prior work on AGR and gender diversity has focused on academic AGR literature, we provide an in-depth analysis of the infrastructure that supports existing commercial systems already widely available for third party use. Our research demonstrates how gender is conceptualized throughout multiple layers of FA systems, including an analysis of both classification and labeling functionality. We enumerate what options are currently available to third party clients, providing insight into the implications of the underlying infrastructure. Finally, we detail ethical decisions we made that may be of benefit to scholars conducting similar research—particularly as it pertains to minimizing misuse of user data when working with cloud based services. The findings presented in this paper can lower barriers for stakeholders evaluating blackbox systems, support current approaches to fairness research, and open doors to more creative ways of imagining gender in algorithmic classification systems.

2 RELATED WORK

We situate our research within three areas of scholarly work in social computing. We start by discussing different approaches to studying identity, specifically social identity and technical identity. Prior identity scholarship grounds our approach to gender as a fluid social construct that we argue is being encapsulated by technical systems. Next, we outline previous work on infrastructure, exhibiting how past scholars have examined identity in the context of infrastructure—and also the various approaches infrastructure research has taken. We conclude by detailing prior work on facial analysis technologies, with a specific focus on gender classification.

2.1 A Roadmap to Algorithmic Identity: Navigating Social and Technical Identity

Human identity is complex. Scholars and theorists from diverse fields define identity in different, sometimes divergent ways. Identity can refer to an individual's personal identity, their social identity, their professional identity, or their cultural identity. Governments often characterize identity as measurable, as in demography, the sociological study of populations [107]—though this too varies across different nations and cultures. Others focus on the multiplicity of ways complex human identities form, as seen in Erikson's theory of psychosocial development [37] or in Marcia's identity status theory [70]. Other theorists focus on inter-subjective aspects of identity. For example, Judith Butler posited that gender and sex are socially constructed, upheld through culturally temporal beliefs about how gender is and ought to be performed [21]. Gender is not a fixed or biological category; rather, it is a theatrical act rooted in constructed "social regulation and control" [21]. Similarly, trans scholars such as Jack Halberstam [45] have explored the nuances of masculinity and its construction in society.

Identity has played a critical role in social computing research. Two differing perspectives of socio-technical identity have emerged in HCI and CSCW research: *social identity* and *technical identity*. Social identity work has focused largely on the experiences users have when interacting with technologies; gender has been a critical focus in this work. For example, Ammari et al. explored the performance of fatherhood in online do-it-yourself communities [8]. Haimson et al. examined the practices of disclosure trans users engaged in on Facebook, and the stress associated with it [43]. Similarly, Scheuerman et al. investigated how trans users navigate safe and unsafe social spaces online [98]. On the other hand, technical identity research concentrates on how identity is represented through system affordances—or, as Leavitt defines in [61], the "technical implementation of an individual's presence within a sociotechnical platform." Leavitt explored the concept of temporary technical identities by examining the practices of Reddit users who make temporary accounts for the explicit purpose of later abandoning them. Schlesinger et al. inspected how the ephemeral, anonymous, and localized nature of technical identities on YikYak impacted user and community identity [99]. Finally, Brubaker and Hayes examined the different uses of persistent identities on Facebook vs. "single-use" identities on Craigslist, documenting how system representations supported social interactions, representing user relationships through differing affordances [18].

In this paper, we examine how human identity—and specifically gender—is databased, parsed, and operationalized by algorithmic methods. While scholars have argued for the importance of considering gender in HCI[10, 92], our focus is specifically on the way that gender is represented and produced through data and infrastructure. Our focus on identity draws from humanist scholarship, including Poster's concept of a "data double" (how the self is reduced to simplistic data fields) [87], Critical Art Ensemble's "data body" (how the body becomes tied to endless data collected in service to corporations and the state) [25], and Cheney-Lippold's concept of a "new algorithmic identity" (how algorithms determine identity from anonymous trace data) [23]. We discuss how automated

facial analysis technology merges the social and technical identities of individuals using their physical appearance—into a *new* new algorithmic identity. We posit that this form of algorithmic identity blurs the social and the technical perspectives of identity, calcifying social identities into fixed, technical infrastructures. Through an examination of our findings, we discuss how this algorithmic identity might affect the way an individual might interact and function within society.

2.2 Coded Categorization: Classifying and Infrastructuring Identity

Categorizing, classifying, and databasing the complexity that is human identity into information systems is laborious and often muddled. In their work, Bowker and Star highlight the cultural, political, and historical decisions underlying the creation of classifications and standards, introducing new methods for examining the infrastructures of information systems [85]. Since then, numerous social computing researchers have inspected the architectures of classification. For example, Blackwell et al. describe the marginalization of users whose experiences with harassment are invalidated when they meet rigid classifications [16]. Feinberg et al. have employed a critical design perspective to privilege the "others" that fall between categories in database infrastructures [38]. Harrell adopted this perspective to examine stereotypes and stigmas reified in games and social media websites [47]. In the realm of algorithmic fairness, Obermeyer and Mullainathan uncovered significant racial bias against black patients in algorithmic auto-enrollment for care programs [82]. Benthall and Haynes proposed an unsupervised training method for detecting racial segregation and using that information to mitigate racial biases in machine learning systems [12].

Examining classification infrastructure has provided a particularly useful lens for gender research. Scholars have long been critiquing the use of gender classification in information infrastructure, as the use of gender on state-mandated forms of identification (e.g. [30, 79]) and sex-segregated spaces, enforced through classificatory signage and expectations around gendered presentations (e.g. [11, 110]). Social scientists Currah and Mulqueen examine the TSA and airport security as a case study for how gender presentation is an unreliable metric for classification [26]. Davis draws on historical cases of gender discrimination to question the purpose of gender classification as a whole, arguing that these classifications harm cisgender[2] individuals as much as they do trans ones.

Social computing researchers have extended this conversation to include social media platforms and other computational systems. Haimson and Hoffman critiqued the underlying structures Facebook has utilized for enforcing authenticity on its platform, disproportionately impacting trans individuals and drag performers [44]. Bivens and Haimson discuss how, even after Facebook opened its user profile up to increased gender flexibility, it has continued to fit non-binary genders into binary classifications to serve advertisers [14, 15]. Game scholar Dym points out the constraints of gender portrayals in games leading to fan portrayals that expand these embedded limitations—and even in fandom portrayals, the tagging structures of fandom communities still limit fans' more expansive approaches to gender [33].

When considering the role of infrastructure in HCI and interface design, Edwards et al. contribute a useful lens highlighting the layers of code, toolkits, and libraries [34]. They argue for engaging more deeply with infrastructural layers in addressing user experience concerns, which ultimately constrain the uses of a system. In this work, we specifically study computer vision infrastructure, highlighting how it could impact third-party applications and individual end-users. We synthesize this approach using two key concepts from Bowker and Star [85]. The first, *torque*, describes how classification systems introduce tension into the lives of the individuals being classified—when "the 'time' of the body and of [its] multiple identities cannot be aligned with the 'time' of the classification system (page 190)." The second concept, *residuality*, describes the "other" categories

[2]an individual whose gender identity aligns with the one assigned at birth. An abbreviated version is "cis."

that do not make it into the classification system; the residual represents the metaphorical leakage as the othering of individuals breaks down. Our study explores how simple classifications found in facial analysis and image labeling technologies, and their applications, might torque individuals whose lived experiences are otherwise residual, or invisible, to the system.

2.3 The New Face of Technical Identity: Automated Facial Analysis Technologies

Automated facial analysis is a collection of computer vision tasks for processing and analyzing digital images or videos that contain human faces. Common facial analysis tasks utilize machine learning for facial detection and facial recognition [63]. Both facial detection and facial recognition frequently include methods for facial classification, a probabilistic operation for classifying attributes about human faces (e.g. [62, 66, 117]). Beyond facial detection and recognition, FA can be designed to predict certain details about a face: its age [97], its ethnicity [67], its affect [108], and its gender [80]. This is generally done using training data comprised of large sets of images, which are qualitatively labeled by humans for specific characteristics (e.g. [9]).

Classifying the gender of human faces is one of the most pervasive forms of facial classification. Gender classification, also known as automatic gender recognition (AGR), is commonly available in most commercially available facial analysis services (e.g. [3, 4, 6]); it is also a major focal point for research looking to develop new machine learning techniques or improve the accuracy of known ones (e.g [7, 59, 93]. Automated facial analysis, and its associated gender classification techniques, have been proposed for a variety of applications, such as access control, real-time security, targeted marketing, and personalized human-robot interaction (e.g. [80, 81, 90, 100]).

Facial analysis, and its deployment for identifying fragments of human identity, has quickly emerged as both a field of rapid development and as a site of controversy—amongst fairness, bias, and ethics researchers, and also the public and popular press (e.g. [101, 105]). AGR has also become a site of examination and critique. One major mode of inquiry has been investigating the accuracy of of gender classification. Jung et al. used three different datasets to examine the accuracy of Face++, IBM Bluemix Visual Recognition, AWS Rekognition, and Microsoft Azure Face API; they found that all four services had an accuracy of 0.9 (90%) on a binary "male/female" spectrum [54]. Buolamwini and Gebru created a dataset of 1270 faces, intentionally balanced for skin type and male/female equity [19]. Using this dataset, they evaluated three commercial facial analysis services from Face++, IBM, and Microsoft. They found that commercially available applications tended to gender white men correctly, while they misgendered darker skinned women most often [19]. Muthukumar et al. examined the individual features of an individual which might contribute to the high misclassification of darker skinned women, uncovering that neither skin type and hair length are predictive of binary gender classification accuracy [76]. Rather, they found that eye, cheek, and lip features were most predictive of gender, unveiling concerning implications about gender stereotypes related to lip and eye makeup [ibid].

Others have scrutinized the ubiquitous use of a male/female gender binary in AGR. Through semi-structured interviews with trans individuals, Hamidi et al. identified universal concerns with how binary AGR might impact the safety, liberty, and emotional wellbeing of people with diverse genders and gender presentations [46]. Keyes employed a content analysis of academic AGR literature to discuss how it inherently excludes trans people, particularly due to its reliance on physiological characteristics to determine gender [57]. Both of these studies highlight the potential risks AGR has for populations who are already widely marginalized.

We augment previous literature by contributing empirical data on both commercial AGR infrastructures and performance rates on diverse gender images. We compliment this with a qualitative analysis of Instagram user commentary about self-held gender identity to offer new insights into how state-of-the-art facial analysis technology interfaces with gender performativity. We also

extend beyond the analysis of facial images to include a focused examination of the infrastructure tying gender to the larger facial analysis system.

3 PHASE I: TECHNICAL ANALYSIS OF FACIAL ANALYSIS AND IMAGE LABELING

Our investigation began with an in-depth technical analysis of commercially available computer vision services that included facial analysis and image labeling functionality. We start by detailing how these services function, paying close attention to what forms of data are provided and how they are organized. In line with Edwards et al. [34], we argue that these services provide an infrastructure that empowers system designers and developers that make use of said infrastructure, but also constrains the possibilities for their designs. Specifically, Edwards et al. call attention to *interjected abstractions*—the risk of low-level infrastructural concepts becoming part of the interface presented to end-users [34]. We argue that this can occur, but that these abstractions can also be uncritically adopted by developers of third-party applications as well. Developers working with computer vision services may accept the APIs and data produced by these services as representative of the actual world (cf. [18, 115]). Even as computer vision services provide tools that empower designers and developers to create new applications, the data provided by these services also represent a set of affordances that designers can use, naturalizing categories and specific data values.

To select a set of services to study, we reviewed several dozen commercially available computer vision services. We initially identified services to review based on our existing knowledge, previous scholarship on these services, and online articles comparing providers. We compared these services using information from their public facing websites, including advertising and promotional content, technical documentation, tutorials, and demos. We eliminated services that (1) did not classify attributes about a human face or body and (2) did not have publicly available demos to test. This process helped us narrow our list down to the ten services we studied during this phase (see Table 1).

3.1 Functionality of Facial Analysis and Image Labeling Services

The computer vision services we analyzed bundled their features in different ways, but the features we analyzed can be broadly understood as falling into two categories—facial analysis and image labeling.

- *Facial analysis* employs specific feature detection functionality trained for faces. Notable for our current analysis, most services bundled a pre-determined set of classifiers that were not solely focused on facial recognition. Services typically classified and categorized the image relative to other concepts (e.g., gender, age, etc.).
- *Image labeling* (or "tagging" on some platforms) provides a set of labels for objects detected in the image (e.g., young lady (heroine), soul patch facial hair). In contrast to the consistent data schema provided by FA, the specific labels and how many are included varies, depending on what was detected in the image.

To better understand the features of each of these ten services, we used free stock images including people, animals, inanimate objects, and scenery. We included images other than people at this stage to identify differences in the data returned for human versus non-human images. Our analysis of the data returned from each service focused on two levels: the schema of the response and the range of values contained in that schema. To this end, we analyzed technical documentation, marketing materials, and the results from our own tests with these services.

Facial Analysis and Image Labeling Services				
Name	***Service Name***	***HQ***	***Gender Class. Terms***	***Prob. Score***
Amazon	Rekognition	United States	Male/Female	Incl.
Baseapp	DeepSight	Germany	Male/Female	Not Incl.
Betaface	Betaface API	India	Male/Female	Incl.
Clarifai		United States	Masculine/Feminine	Incl.
Face++		China	Male/Female	Not Incl.
Google	Cloud Vision	United States	N/A	N/A
IBM	Watson Visual Recognition	United States	Male/Female	Incl.
Imagga		Bulgaria	N/A	N/A
Kairos		United States	M/F	Incl.
Microsoft	Azure	United States	Male/Female	Not Incl.

Table 1. The set of facial analysis and image labeling companies (and their service name, if it is different) whose documentation we analyzed. The "Gender Classifier Terms" column represents the language used to describe gender classification in the service. The "Probability Score" column indicates whether the gender classifier includes a probability score. Bolded names represent the services we studied during Phase II (see Section 4.3).

3.2 The Schema of a "Face"

The schemas for facial analysis services were elaborate (see Figs. 1-2), but highly varied. Some services provide robust detection of the location of facial "landmarks" (e.g., eyes, nose, mouth, etc.) and orientation of the face within the image (i.e., roll, yaw, and pitch). While facial features may be indicative of gender (e.g., facial morphology [68, 90]), in this analysis, we focus on the classification data that would most commonly be used by third-party developers making use of these services.

All services, save for Google's, included gender and age classification in their facial analysis. We found that services from large tech companies in the United States (such as Amazon Rekognition, Google Cloud Vision, and IBM Watson Visual Recognition) omitted ethnicity and race. However, smaller, independent companies (such as Clarifai and Kairos) and non-US companies (like Chinese-based Face++ and German-based Beta Face) included ethnicity and race.

Finally, some form of "safe search" classification was common across FA services. These classifiers included ratings for attributes like "raciness" or "nsfw." IBM's Watson, for example, includes classification results for *explicit*. Microsoft Azure has two classifiers for *adult* and *racy* content.

When returning classifier results, some services also included a probability score for the result. This was variously referred to as a "score," "confidence score," "accuracy," and so on, but represented the likelihood of the returned value being accurate.

Returning to gender classification, the range of values we observed is important. Gender was defined as a binary—and never a spectrum—with only two categories. Despite often being called "gender," the categories always used the biologically essentialist[3] terms "male" and "female" (as opposed to actual gender identities like "man" and "woman"). One interesting exception was Clarifai, whose gender classifier is specifically termed "gender appearance" and returns the values "masculine" and "feminine." These terms insinuate a potential shift in gender classification from a biologically-associated category that someone is assigned to a perceived quality someone could define.

[3]Focusing on specific, fixed biological features to differentiate men from women.
[4]Original photo attribution: Aiony Haust. Photos have been edited by the authors for the purposes of demonstration.

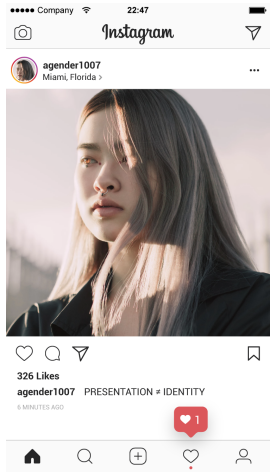

Fig. 1. An example of an #agender Instagram post.[4]

```
...
"age": {
    "min": 20,
    "max": 23,
    "score": 0.923144
},
"face_location": {
    "height": 494,
    "width": 428,
    "left": 327,
    "top": 212
},
"gender": {
    "gender": "FEMALE",
    "gender_label": "female",
    "score": 0.9998667
}
```

Fig. 2. An example of the gender classification portion of the result from IBM Watson's facial analysis service.

```
{
  "class": "woman",
  "score": 0.813,
  "type_hierarchy": "/person
  /female/woman"
},
{
  "class": "person",
  "score": 0.806
},
{
  "class": "young lady (heroine)",
  "score": 0.504,
  "type_hierarchy": "/person/female
  /woman/young lady (heroine)"
}
...
```

Fig. 3. An example of a post in our dataset and its output when run through the IBM facial analysis API.

During our analysis we also noted that probability scores (when included) never fell below 0.5 for the classified gender. Some services, like Kairos, provided two probability scores that total to 1.0 (e.g., "femaleConfidence": 0.00001, "maleConfidence": 0.99999, type: "M"), clearly exposing a binary classifier for which male and female are opposites.

In our initial exploration, we found that the results of gender classification were inconsistent across platforms. However, it was often difficult to determine why. In one instance, for example, a photo of a man dressed in drag from our dataset was classified as female (Watson) and male (Azure). These inconsistencies demonstrate the differences in how gender is operationalized across these services. However, it is unclear whether this is a result of differences in training data and the

creation of models, or if there are more fundamental things at play. We return to this concern in section 4.3 with a more rigorous evaluation.

The uniform simplification of gender across most services was surprising, but also prompted us to consider other places in which gender might exist, but be less structured. With this in mind, we also analyzed the more open-ended image labeling features.

3.3 Labeling

In contrast with the consistent set of classifier data accompanying FA results, the data returned for image labeling is far more open-ended. As previously mentioned, labeling requests produced a list of the objects detected—but the range of labels provided is extensive. While we could not find exact numbers, Google and IBM both claim that they have classification for "thousands" of "classes." As with facial analysis, some services provided probability scores for their results, while others did not.

The absence of an explicit gender classifier, however, does not mean that gender was absent from labeling results. In fact, gender was evident throughout labels, including terms like *"woman," "man," "boy,"* and *"aunt,"*, to name a few. Moreover, unlike the binary gender classification, labels are not mutually exclusive. As a result, we frequently saw images labeled with multiple and seemingly contradictory terms. For example, a set of labels including *"person," "boy," "daughter,"* and *"son"* was not unusual.

3.4 Independence in Classification Tasks

Given FA and labeling were offered by the same provider, we expected there to be consistencies across gender classification in both FA and the gendered labels. If anything, we found the opposite. As evidenced by the prevalence of multiple gender labels being assigned to a single image, gendered labels were decoupled from the gender classifications in FA. Probability scores for FA gender classifications and label classifications would often differ, sometimes greatly. For example, Amazon assigned the label *"female"* (.612) to an image, yet the probability score for this label was much lower than the probability for its female gender classification (.992). This suggests that the gender classifiers in FA services are decoupled from the classifiers used to label images.

These inconsistencies provide insights into how we might better understand probability scores. Clarifai, in particular, provided a unique glimpse into how it discretely classifies gender into two categories by showing the probability scores for both male and female classification. For some of the images we tested, the probability scores were close to one another on both sides of the binary, showcasing *a lack* of confidence in its gender classification. For example, an image of a young woman standing next to a statue was viewed as only .500002 likely to be female, and .49997 likely to be male, tipping the scales towards female just slightly. However, this image was labeled with a series of seemingly contrasting binaries: *"child"* (.986) and *"adult"* (.893), *"boy"* (.91) and *"girl"* (.904). As evidenced by the independent probability scores for these labels, classification is based on the detection of the label or not (e.g., "woman" or "not woman") rather than the either-or binaries (e.g., "female" or "male") we observed with gender classification in facial analysis.

3.5 Takeaways from Phase I

During our initial analysis, we found inconsistencies across services in how gender was classified. Despite these inconsistencies, however, both facial analysis and image labeling used binary language to describe gender. Moreover, the inconsistencies between how gender and labels were classified, even within the same service, suggests that classifiers are developed independently from each other and, subsequently, gender is being operationalized in a piecemeal fashion.

Examining the APIs associated with these services, as if we were third-party developers, impressed the important role of data schemas. The structure of the data returned by gender classifiers to third-party developers presents gender as a property that can be easily incorporated into the design of their applications. In the process, however, much of the context around gender, or even a classifier's certainty, can be lost. Established literature on algorithmic fairness has noted how biased training data can result in biased classification and recommendations. What we see here is how these systems can also produce bias in how they framed their results—in this case, through the schemas around which services and their APIs are constructed.

What our initial study did not tell us is how image classification performed on a diverse set of images. Moreover, our analysis of image labels was not exhaustive. In the next phase of our research, we sought to quantify the performance of gender classification across a diverse set of genders and analyze a larger and more diverse set of image labels.

4 PHASE II: EVALUATING FIVE FACIAL ANALYSIS AND IMAGE LABELING SERVICES

Having detailed how FA and image labeling services function, and the affordances that they provide to third-party developers, we analyzed how a subset of these services performed with a dataset of gender diverse images. Specifically, we:

(1) Performed a system analysis to understand the affordances of these systems.
(2) Manually determined the performance of the gender classifiers found in facial analysis services using a dataset of diverse gender images.
(3) Conducted a qualitative error analysis to understand how the gendered language of the system compares with the gender expressions of the Instagram dataset authors.

We start by describing how we narrowed the services we studied. We then outline the construction of our dataset, including how we selected the genders we studied and the inclusion criteria we used for images. For both of these, we discuss how ethical considerations impacted the methodological choices we made.

4.1 Facial Analysis Services

To conduct a more in-depth analysis, we selected five services based on two criteria: (1) a diversity of classification and labeling affordances (in other words, the types of information returned about images) and (2) their presence in the market and size of their user base. Specifically, Amazon [1], Google [5], IBM [6], and Microsoft [3] have been notably active this past year in their investments and involvements in the state of facial recognition technology (e.g. [36, 69, 78, 88]). We also included Clarifai [2], a small startup company known in AI and facial recognition markets for its involvement in government contracts [75]. These five services each included gender information in either their facial analysis feature, their labeling feature, or both.

4.1.1 Ethical Considerations for Service Selection. Following our technical analysis of ten services (see section 3), we also decided to eliminate several services from our dataset evaluation on ethical grounds. It is common for service providers to make use of the data provided to them for product development. However, such data use and retention policies present ethical concerns in the context of this study. Face++, for example, retains the right to use analyzed photos for internal research and to improve their products. We decided against using these services because we were unsure what unethical or potentially harmful use cases these services might then use our data for. We reviewed the terms of service (TOS) for each service to ensure that the service we used did not store images for any purpose or use the images to further train models, or alternately, allowed us to opt-out of data storage and training.

Folksonomies Turned Hashtag			
	Instagram #		***Instagram #***
AFAB	28,191	***Man***	36,466,751
Agender	1,864,879	*Neutrois*	28,060
AMAB	20,332	***Non-Binary***	2,780,477
Androgyne	223,125	*Pangender*	156,596
Bigender	855,370	*Polygender*	103,620
Cisgender	97,971	*Third Gender*	12,592
Demiboy	597,320	*Trans*	5,933,800
Demigirl	592,703	*Trans Feminine*	26,060
Female	6,379,367	***Trans Man***	843,139
Femme	3,132,240	*Trans Masculine*	132,380
Gender Nonconforming	84,780	***Trans Woman***	452,743
Genderless	236,082	*Transgender*	7,849,435
Genderqueer	1,990,117	*Trigender*	171,539
Male	6,884,437	***Woman***	41,269,789

Table 2. The folksonomies generated by the seven author contacts. The Instagram Posts column indicates the number of posts under the associated hashtag. The bolded gender labels indicate which genders we chose to use for our final dataset.

4.2 Dataset

Next, we constructed a dataset of gender diverse images. After sampling and cleaning the data, our dataset comprised of 2450 photos that included a face, were posted publicly on Instagram, and were labeled with a gender hashtag by their author. Our dataset contained seven different genders, with 350 images for each.

To identify a diverse set of gender hashtags, we crowd-sourced gender labels from seven author contacts, all of whom were queer, trans, and/or non-binary individuals. This method was especially useful for generating a list of genders beyond cisgender and binary ones, as trans communities often develop and employ folksonomies to self-identify [28].

Our contacts provided a total of 24 unique gender folksonomies. From these, we selected seven diverse genders, weighing what was most commonly used on Instagram, while also excluding folksonomies that could have multiple meanings (e.g., androgynous, queer, transgender). The final folksonomies-turned-hashtags were #man, #woman, #nonbinary, #genderqueer, #transman, #transwoman, and #agender (see Table 2).

Having selected a set of gender hashtags, we then used an open source Python tool[5] to collect a sample of photos and their associated metadata for each hashtag. To ensure a diversity of images, we collected images from Instagram's Recent feed rather than Instagram's Top feed. (The Top feed is algorithmically curated to showcase popular content and is biased towards celebrities, influencers, and other popular accounts.) Examining the photos returned, we discovered that a number of the images were irrelevant for our analysis, such as memes or illustrations, or not suitable for simple facial analysis (e.g., images which are pixelated, low-quality, distorted by filters, etc.). As such, we adopted the following three inclusion criteria for photos:

(1) A single human face must be present. Images without a face, or with multiple faces, were excluded. Including images with multiple faces would make it difficult to differentiate which data is associated with which face in the image.

[5] https://github.com/rarcega/instagram-scraper

(2) 75% or more of the face must be visible. We eliminated faces that were cropped out of the photo or hidden behind an object, hand, or hair.
(3) The image must be clear and not visibly altered or filtered. We eliminated photos where an individual's face was unclear or heavily distorted with filters; we also eliminated images that were low-quality or pixelated.

Knowing that facial analysis technologies are not 100% accurate [42], removing these photos allowed us to focus on gender presentation in ideal circumstances and better isolate how these services classify and label gender in each individual image.

Our final dataset consisted of 2450 total photos associated with the seven hashtags, with 350 images each (a breakdown can be seen in Table 2). After finalizing our dataset, we processed all images through the five selected services. We used the results in three ways. First, we performed a quantitative evaluation of the gender classification returned by face analysis. Next, we qualitatively analyzed results from labeling requests. Finally, to understand how self-held gender aligns with computer vision classifications, we conducted a content analysis of a subset of Instagram captions to compare with the face classification and labels.

4.2.1 Gender Hashtag Nuances. The fluid and imbricated nature of gender makes it difficult to divvy it up into neatly divided hashtags. In fact, many of the posts we collected used two or more of these hashtags. For this reason, explaining the nuanced meaning behind these labels and the way we employ them for this study is necessary.

Foremost, even the gender binary is not easily split into simple cisgender or transgender categories. For example, #man and #transman could easily represent the same person: like cis men, many trans men identify with the label "man" without the trans prefix. The same is true with trans women. Thus, we cannot assume that the men in #man and the women in #woman are cisgender. In the same way, we cannot assume that the trans men in #transman and the trans women in #transwoman identify solely with the binary. For example, some trans women may identify with the label trans women, but not women. Non-binary, meanwhile, is often used as an umbrella term. #agender and #genderqueer may fall under #nonbinary; individuals who tag with #transwoman and #transman may also be non-binary.

For our quantitative analysis (see section 4.3), we collapsed gender into binary categories for the purpose of assessing performance of binary gender classifiers. #man and #transman became ground truth for "male" and #woman and #transwoman became ground truth for "female." This allowed us to understand how binary gender classification performed on binary genders, whether cisgender or trans.

While we acknowledge the issues with collapsing genders in this way, doing so allowed us to assess and compare the true positive rates for each gender category. However, for our qualitative analysis, we examine the nuance of gender in comparison to the binary outputs of facial analysis services. We engage with how users self-describe their own genders in their Instagram photos and compare these with how facial analysis and image labeling services classify gender. This surfaces how classification binaries fail to capture the full range of gender.

4.2.2 Ethical Considerations for Data Collection and Dataset Construction. We recognize the sensitive nature of collecting public user data for research purposes. Thus, we considered what the benefits and the risks were to choosing this method [39]. We feel that this work is important in highlighting the current limitations, and potentially negative implications, that FA and image labeling technologies have for individuals of diverse genders. Due to the lack of available ground truth image data of trans individuals, we felt it was important to work with a ground truth dataset that did not contradict users' self-held gender autonomy. We did not collect or store Instagram

usernames. Furthermore, we destroyed all of the files containing the images and posts after the completion of our analysis. The dataset we constructed will not be published, to both protect user identity and to ensure user images are not appropriated for unethical or harmful research.

To further protect the identities of the Instagram users in our dataset, the images we include throughout this paper are not part of our dataset and serve only as exemplars. Examplars are images from Unsplash, a stock website that provides license for unlimited image use for commercial and noncommercial purposes—they do not represent true ground-truth data. We also paraphrased or created composites of user quotes, rather than directly quoting users, so that the identities of users cannot be identified through search [71]. We believe that the steps we have taken mitigate the possibility of harm to users to such a degree that the benefits outweigh the risks.

TPR Performance Per Gender Hashtag													
Hashtag	***Amazon***			***Clarifai***			***IBM***			***Microsoft***			***All***
	T	*F*	*TPR*	*T*	*F*	*TPR*	*T*	*F*	*TPR*	*T*	*F*	*TPR*	*Avg*
#woman	348	2	99.4%	333	17	95.1%	345	5	98.6%	100	0	100.0%	98.3%
#man	334	16	95.4%	344	6	98.3%	341	9	97.4%	348	2	99.4%	97.6%
#transwoman	317	33	90.6%	271	79	77.4%	330	20	94.3%	305	45	87.1%	87.3%
#transman	216	134	61.7%	266	84	76.0%	250	100	71.4%	255	95	72.8%	70.5%
#agender, #genderqueer, #nonbinary	—	—	—	—	—	—	—	—	—	—	—	—	—

Table 3. The True Positive Rate (TPR) for each gender across the face calls of each of the facial analysis services we analyzed.

4.3 Performance of Facial Analysis Services on a Diverse Gender Dataset

After completing our system analysis (see Section 3), we sought to understand how FA and image labeling services performed on an image dataset of people with diverse self-identified genders. In this section, we specifically focus on the gender classification provided in the results of facial classification requests.

Using the gender hashtag provided by individuals in their Instagram post as ground-truth data, we calculated the accuracy of gender classification results from four services across 2450 images. For this analysis, we examined results from Amazon, Clarifai, IBM, and Microsoft. We excluded Google as its Vision service does not provide gender classification.

We calculated the True Positive Rate (TPR) (also called recall) for each gender hashtag across each service. For the purposes of this study, we refer to this as "accuracy"—the accuracy at which the classification correctly identified the ground truth gender of the person in the image. Finally, it is important to note that we analytically calculated the accuracy rate for #agender, #genderqueer, and #nonbinary as 0%. As noted in section 3, FA services with gender classification only return binary gender labels. Given that these three genders do not fit into binary gender labels, it is not possible for any of the services we evaluated to return a correct classification.

Our analysis reveals differences across both genders and FA services (see Table 3). Differences in true positive accuracy likely indicate how these models are trained to recognize contrasting "female" and "male" features (e.g., lips and cheekbones [76]). Due to the stark differences in accuracy between

cisnormative images (#man and #woman) and trans images (#transwoman and #transman), it is likely that the training data used to train FA services does not include transgender individuals—at least those who do not perform gender in a cisnormative manner. Differences between services make it evident that each service not only employs different training data to classify gender, but potentially different requirements for the underlying infrastructure driving the task of gender classification. These differences result in subjective notions about what male and female actually are to the respective system.

As seen in Table 3, #woman images had the highest TPR rate across all services, with the exception of Clarifai, which classified men more accurately than women. #woman was classified correctly, on average, 98.3% of the time, with Microsoft providing the highest TPR rate and Clarifai the lowest. #man had the second highest TPR rate, also with the exception of Clarifai. The average correct classification rate for #man was 97.6%. Microsoft, again, correctly classified the greatest number of images and Amazon correctly classified the least number of images. These high rates of true positive accuracy suggest that the training data used to train "male" and "female" classification aligned with cisnormative gender presentation. By extension, it may be the case that service providers considered cisnormative images best suited to the task of gender recognition when creating training datasets, technologically reproducing gender binaries in these systems.

When comparing #transwoman and #transman with female and male classification, respectively, true positive accuracy decreased—particularly for #transman. Images from the #transwoman dataset had the second lowest TPR rates across all services, averaging at 87.3%. IBM had the highest accuracy of #transwoman images, while Clarifai had the lowest accuracy; the difference between the two was 16.9%. These differences in classification accuracy suggest that these services are using different training data to classify what "female" looks like. The #transman dataset had the lowest true positive accuracy, averaging at 70.5%. Microsoft had the highest accuracy (72.8%), while Amazon had the lowest (61.7%). The difference between these two services was 11.1%. While, in general, accuracy rates for #transman were poor, the difference in accuracy across services similarly suggests differences in the range of images being used to train classifiers as to what constitutes "male." The training data selected likely excluded non-normative gender presentation to a higher degree than that found in "female." In other words, cisnormative masculine presentation was likely less varied and diverse in the training data. TPR differences—for men and trans men, and for women and trans women—suggest a subjective view, on the part of computer vision services, as to what "male" and "female" is.

In comparing the two trans datasets, we illustrate the differences across binary trans accuracy rates. #transwoman images had a higher accuracy rate across all services than #transman images. The most stark difference between #transwoman and #transman rates was within Amazon, with a difference of 28.9%. The lower TPR for #transman across all services suggests that variance of images for "male" is smaller, resulting in services that understand "male" as something more specific and bounded than "female." The two genders with the greatest TPR rate difference were #woman and #transman. The high TPR of #woman further suggests either a greater range of "female" gender presentations in the training data used by these providers or a more normative gender presentation on the part of the trans women in our dataset. The greatest difference here was again found on Amazon, which classified #woman accurately 37.7% more frequently than #transman. This also suggests that models trained to recognize "female" images are recognizing images of trans men as female as well.

To understand the different range of images classified as male or female, we qualitatively examined misclassified images. As suggested by the differences in classification rates, the specific images that

[5]The assumption that all individuals are cisgender, and thus cisgender is the expected norm.

services misclassified varied. Clarifai misclassified the greatest number of #woman images. For instance, it was the only service to misclassify an image of a woman with a ponytail wearing a leather jacket, yet it was highly confident in its classification (.951). Discrepancies between what images were misclassified across services again suggests differences between datasets used to train these systems. However, at underlying all of these systems is a design choice about how to operationalize gender in the first place, and thus what data should be used to train these classifiers.

These findings led to more questions about how gender is metaphorically "seen" and classified by computer vision services, prompting us to conduct a qualitative analysis of the *labeling* assigned across diverse genders.

4.4 How Gender is Labeled, How Labels are Gendered

While most computer vision services include gender classification in their facial analysis results, many also include image labeling. While standard practices in computer vision and machine learning suggest that gender classification and image labeling are algorithmically distinct, we felt that studying the labels was important in presenting a holistic view of how gender understood within these services. We qualitatively analyzed the labels assigned to Instagram posts across the five services. We examined the associated computer vision labels using a subset of 100 Instagram posts per gender hashtag from our full dataset. We hand-coded these labels for gender-specific concepts, including explicit gender labels (e.g. *"woman"*) and implicit gendered concepts (e.g. clothing types, aesthetic qualities).

Our analysis cannot identify how these labels were derived. However, typical industry practices involve creating lists of thousands of concepts that these systems can detect in a fairly haphazard fashion. When approached as a technical problem, designers may choose to focus on the ability to accurately detect the given object without considering the social meaning of such objects or the potentially harmful implications.

In this section, we outline how images were labeled *within* and *across* services. We demonstrate the relationships between these outputs and the Instagram posts by describing the people in the images. We also present the probability scores associated with each label in our presentation of these findings, when services made them available.

4.4.1 The "Cultural" Language of Labels: What is Feminine and What is Masculine. Many images were unanimously associated with explicitly gendered labels. For example, a photo of a blonde woman in a gown was labeled *"woman," "girl," and "lady"* by Microsoft. Labels were often redundant in this way, assigning many different variations of *"woman"* to single images. It was also common for labels to be feminized versions of otherwise gender neutral words, like *"actress"* and *"starlet."* This was rare for male-classified photos. The only examples we found were from IBM: *"muscle man" and "male person."*

Gender was often implicitly manifest in the labels as well. For example, a portrait of a woman with long dark hair and heavy makeup was labeled by Microsoft as *"beautiful"* and *"pretty,"* concepts often associated with feminine women. Like Microsoft, Clarifai also labeled this image with the traditionally feminine label *"beautiful"* (.937). In contrast, labels like *"beauty"* and *"pretty"* were rarely assigned to male-classified images. Likewise, it was uncommon for traditionally masculine labels to be assigned to women. For example, none of the labels contained concepts like *"handsome"* or *"rugged."*

Gendered labels extended to specific physical features and clothing as well. Services recognized traditionally masculine facial features, like beards, moustaches, and stubble. In our analysis, we did not see traditionally female equivalents. However, feminine gender was assigned to garments quite

often. For example, IBM returned labels that explicitly included femininity: *"halter (women's top),"* *"women's shorts"* and *"decolletage (of women's dress),"* for example.

The implicit and explicit gender in the labels we observed showcases how the services we studied conceptualize binary genders—what is female and what is male, what is feminine and what is masculine. Masculinity was often portrayed as the "neutral" position—it was rarely used as a modifier. Femininity was, however, used to further describe otherwise neutral terms. Labels like *"actress," "heroine," "starlet,"* and *"women's apparel"* invoked an explicit femininity not present in masculine labels.

4.4.2 A Black Box of Gender Labeling. Our analysis is unable to determine why services assign the labels they do to any given image. However, when labels are gendered, it can be assumed that those annotations are based on cultural gender norms, as found in previous facial analysis literature [76, 86]. For example, that women wear makeup and men do not. Connections like this may be responsible for the labels assigned to an image of a #transman with long wavy hair, winged eyeliner, and red lipstick: Microsoft labeled this image with *"woman"* and Google with *"lady"* (.864).

On the other hand, gender classification did not always seem tied to binary standards of gender performance. In one case, a thin blonde "#femme" trans woman with fuschia lipstick was labeled *"boy"* by Microsoft. This seems to contradict the notion that cultural constructs of gender performance, such as makeup being feminine leads to *"female"* labeling. These examples demonstrate that it is actually impossible for human beings to determine how these services are making labeling decisions and what specific objects in an image these labels are tied to. While the cause may be self-evident with a label like *"toy"*, it is less clear with labels such as *"pretty"*.

Increasing the complexity, image labeling does not necessarily consider the person in isolation. Instead, myriad labels are typically produced based on whatever is detected in the image. As a result, there were often seemingly divergent gender labels assigned to the same photograph. As one example, Microsoft labeled and image of a woman with long dark hair and a low cut red dress with *"woman," "girl," "lady,"* as well as *"man."* Likewise, after accurately classifying the gender of a muscular trans man with glasses, Microsoft provided both *"man"* and *"woman"* as labels. While these divergent labels still reinforce a gender binary, they also suggest a lack of clarity on what about these images results in the rather general labels of *"man"* and *"woman"*. Typical industry practices involve creating lists of thousands of concepts that these systems can detect in a fairly haphazard fashion. Given the breadth and variety of concepts that may be included on such a list, designers may focus on accuracy of detection and not the social meaning of such objects.

4.5 Self-Identified Gender versus Computer-Classified Gender

After analyzing computer vision services, their facial classification, and their labeling, we wanted to understand how gender was presented by people in the larger context of their image captions. We performed a content analysis of these captions, specifically focused on how users discussed gender in their original Instagram posts. While #man and #woman users typically did not comment extensively on gender, when analyzing trans hashtags, we identified three categories: personal narratives, gender declarations, and critical commentaries of the relationship between self-presentation and gender. While the results of classification and labeling may be presented to end users in different ways in third-party applications, and some of these results may be entirely invisible to them, we present each of these in more detail as a way of illustrating the potential impact gender classification infrastructure could have on real people. By doing this, we explicitly build on prior work examining the potential harms FA might cause in real-world scenarios [46, 57]

4.5.1 Personal Narratives Highlight Potentials for Affirmation and for Harm. It is possible that facial analysis technologies could be affirming to trans users with binary genders, as has also

been discussed by participants in [46]. Many users expressed struggling with feelings of gender dysphoria, the emotional distress associated with an individual's experience with their gendered body or social experiences. In another photo, a #transwoman user posted a smiling selfie. This user also wrote that she was having a hard day because *"dysphoria is driving [her] nuts."*[6] The services assigned this image female-centric labels, like Microsoft's *"lady"* and Clarifai's *"girl"* (.981). In these cases, classifiers categorizing her as female in might be affirming.

Of course, while labeling might be affirming to some binary trans individuals experiencing dysphoria, the high rates of misclassification also presents the potential for *increased* harm. The impact of these misclassifications can also be connoted when comparing them with user statements about gender dysphoria and misgendering. For example, Microsoft, IBM, and Clarifai misgendered a #transwoman who lamented on her post: *"I spent an hour getting ready just to be be addressed as 'sir' at the store."* In another example, a #transman expressing *"severe dysphoria"* was misclassified as female by the all of the services. Effectively, system misclassification could have the same negative impact as human misclassification, or even compound everyday experiences of misgendering.

Multiple gender labels could also be problematic when classifying binary trans individuals. A trans woman who wrote *"I've felt dysphoric the past few days,"* was also gendered female by all of the services, but was labeled *"woman," "girl,"* and *"man"* (Microsoft). The presence of a "male" gender label, which cannot be associated with specific characteristics, might also hold negative connotations for those dealing with gender dysphoria.

4.5.2 Classifications Can Never See Non-Binary Genders. The impact of incorrect classification for non-binary people is evident in many of the captions we examined. For example, a #genderqueer user wrote in all caps: *"THIS IS A NON-BINARY ZONE. DO NOT USE GIRL/SHE/HER/HERS."* This user was then classified as female by every service. As described in sections (3 and 4.3), there were no classifications outside of male or female. While gender-neutral labels were provided for some images (e.g. *"person," "human"*), gender was always present in the facial analysis services that used gender classification (every service besides Google)—and also when other explicitly gendered labels were provided.

These binaries also reinforced concerns about gender presentation in non-binary individuals. For example, a #nonbinary user posted a photo of themselves wearing heavy winged eyeliner, but discussed in their post the *"dysphoria"* and *"inner turmoil"* they experience *"when wearing makeup as a non-binary person"* due to makeup's association with femininity. All of the services classified this person as female. Descriptions of "inner turmoil" at being associated with the incorrect gender offer insight into the potential emotional and systemic harms facial analysis and image labeling systems might have when classifying nonbinary genders in real-time.

4.5.3 Declarations of Gender that Cannot be Seen by Computer Vision. These systems lacked the ability to contextualize implicit and explicit visual markers of gender identity, particularly in the context of trans images. For example, Microsoft misclassified an image of a bearded #transman holding up a syringe, presumably for testosterone injection based on the Instagram caption which read: *"I'm back on T (testosterone) after months so hopefully I'll be back to myself."* This is considered a definitive marker of hormone replacement therapy and trans identity, and includes "insider" markers contextual to trans communities. Another photo was of a shirtless #transman with top surgery[7] scars, a visual marker of his trans identity. This image was classified as female by Amazon (though male by every other service).

[6] As described in 4.2.2, this quote is paraphrased to protect user identity.
[7] A term used to describe a gender confirmation procedure resulting in the removal of breasts [113]

As evidenced through these services' abilities to only recognize "male" and "female," they also have the inability to recognize whether someone is transgender, binary or not. So, while some users expressed pride in their identities, the systems are unable to affirm this. For example was of an #agender person wearing a t-shirt that read *"not cis."* The underlying infrastructure would not have the ability to recognize trans from this declaration. Instead, this image was classified as male by Clarifai, IBM, and Microsoft, but female by Amazon. While gender labels that align with user gender expressions could be affirming to their journeys, these services are still unable to recognize their identities as trans.

4.5.4 User Commentary Critiquing Gender Binary. Many users also critiqued the notion of gender being tied to performance or external appearance. When examined alongside the classification infrastuctures highlighted in previous sections, these critiques could be viewed as a point of contention aimed at the premise of technologies like facial analysis and gendered image labeling. For example, a #genderqueer user expressed frustration with the normalization and authority of cisnormative binary gender, writing a series of statements that read: *"Down with cissexism. Down with cisnormativity. Down with cis privilege."*

Another agender user critiqued the historical representation of gender as solely binary in Western cultures. They wrote that *"some transphobics say that people are inventing new genders ... but they aren't new."* They wrote a commentary in their caption outlining the history of two-spirit and transgender roles in Native American life (*cf.*, [106]).As presented in the previous section, the classification and labeling schemas, as well as the higher performance rate on binary genders, privileges the cisnormativity these users are critiquing.

5 DISCUSSION

What is gender? A simple question with no single answer. Gender can be understood through a multitude of perspectives: a subjectively held self-identity [103], a self-presentation to others [41], a social construct defined and maintained through performative acts [21], and a demographic imposed by society [40, 95]. In the context of computer vision, we have shown how the design and use of facial analysis and image labeling systems collapse these perspectives into a singular worldview: presentation equals gender.

Forms of self-presentation are encoded into computational models used to classify these presentations. When classifying gender, designers of the systems we studied chose to use only two predefined demographic gender categories: male and female. As a result, these presentations are recorded, measured, classified, labeled, and databased for future iterations of binary gender classification. These gender classification models are then bundled up for commercial use, often in the form of cloud-based services, providing an infrastructure that third-parties can use to create or augment their own services. In the process, these services propagate a reductionist view of gender provided by the underlying infrastructure. Self-identity is not used by computer vision systems. After all, it cannot be seen.

In order to illustrate how gender is used by computer vision services and experienced by gender diverse individuals, we synthesize our findings through an engagement with Butler's notions of gender performativity ([21]) and Bowker and Star's notions of residuality and torque ([85]). We map our discussion to Edwards et al.'s problem of infrastructure [34], discussing how the perspectives of identity outlined above interact across three layers: the infrastructure, the third-party applications that make use of that infrastructure, and people. Through this discussion, we highlight the layers of translation gender is sifted through as it moves from human being to infrastructure, and then back again.

5.1 Classifying Human Gender Performativity through Infrastructure

Literature on gender classification has highlighted numerous methods for identifying gender (e.g. periocular regions [68]; facial morphology [90]; lips, eyes, and cheeks [76]), but how, when, and by whom gender is embedded into the pipeline of data, labels, and models is opaque to outsiders (e.g. [56, 89]). However, in examining the commercially available affordances and infrastructure, our findings shed light on how the visible presentation of an individual and their performative expressions of gender, through grooming and style, are used by computer vision systems in two ways.

First, through gender classification in facial analysis services, we see binary gender categories applied to individuals. Second, through image labeling, specific aspects of an image are detected and assigned a descriptive label (e.g. beard). The services we studied adopt a particular cultural view of gender that privilege self-presentation and gender performance. However, the manner in which this cultural view relies on presentation can be seen as archaic and normative, adopting systematic demographic gender categories that embrace the binary. We expound on this perspective by unpacking the infrastructure underlying both facial analysis and labeling.

Facial analysis makes use of the most rigid gender categorizations= within commercial computer vision services. Even when the self-expression defies the binary mold, FA employs binary gender classification in a way collapses diverse expression and reinforces what gender should look like. Even if a model is trained to recognize diversity in images of men and women, it can only apply those learned standards in a constrained classification environment when classifying images of trans people. A diversity of training data will not address this bias. Our analysis of FA infrastructure suggests that (with one exception, Google Vision) designers of these services decided to first, include gender in their products, and second, define it as "male" *or* "female." The bias in gender classification cannot be attributed to algorithms alone. Its root sits with how designers conceptualized the problem in the first place.

However, it is important to note that rigid approaches to gender classification is not inherent to all of image classification. We found that, in comparison with face classification features, labeling features had the ability to assess images of people in ways that were gender neutral (e.g. the label "person") or ambiguous (e.g. including multiple gender labels). Examining how labels manifest in images across all genders suggests traditional performative markers of binary genders might not be as inherent to facial classification decision-making as one might expect. Not only can "man" and "woman" exist within one image, labels can represent concepts independent of gender identity: men can wear makeup, women can have beards. Label classification is decoupled from facial classification; they do not impact the results of the gender classifiers we analyzed in facial analysis services. Perhaps this is beneficial, because binary gender classification is not determined by labels for concepts like makeup or beards.

Despite the potential occurrence of multiple labels, gendered labels themselves still typically conformed to binary notions of gender. We saw labels for man, male, boy–but not trans man. Moveover, while labels associated with men were often gender neutral, we found that women were often positioned as an outlier. Many of the labels for feminine presentation used terms that were explicitly gendered synonyms of otherwise gender neutral concepts (e.g. military woman, gown (of women)). Specifically with labels, it is important to emphasize the subtleties in how concepts were gendered—in many cases, even when unnecessary. The abundance of gendered labels we observed points to the importance of considering gender beyond the training of classifiers, but also in the seemingly mundane human work of create labels that will be associated with these classifiers.

Finally, labeling—because it is focused on discrete object detection—does not consider the other objects detected as contextual factors. While in many cases this may contribute to the plurality

of concepts identified in images, this also presents an interesting challenge when classifying self-identity. Many of the posts by trans users included context clues intended to communicate details about an individual's gender to their viewers. Details like wearing a tee shirt with "not cis," wearing makeup as a non-binary person, or writing in a social media profile that you inhabit a "non-binary zone"—is lost when using simplistic object-based classification systems.

There is a bias encoded into systems that render only specific gender performances and specific genders visible. The consequence of current computer vision infrastructure is the erasure of residual categories of gender—categories which cannot exist in a system that is trained to recognize only traditional notions of male and female. As Bowker and Star explain, there is value in exposing residual categories: "[T]hey can signal uncertainty at the level of data collection or interpretation" especially in situations where "more precise deisgnation[s] could give a false impression" of the data. [85, p. 150]

For those who fall into the residual categories of computer vision systems—whose gender cannot be seen by gender classification schemas—the likelihood of experiencing torque is high. An alternative to prescriptive binary gender classification might lie in embracing a polyphony of performative features, embedding labels into the infrastructure with the intention of supporting gender fluidity. Instead of collapsing gender identity to a single category (as occurs with current gender classifiers), computer vision services could embrace the fluidity of gender performativity by providing more comprehensive and inclusive gender concepts in their labeling schemas.

However, supporting a larger number of gender identities and broader definitions of any given gender is not without its limits. Bias goes beyond models and training sets, or even a new approaches that might consider assigning multiple genders to an image. FA is limited by the premise that gender can be seen. Trans scholar Viviane Namaste's critique of gender and queer studies is instructive here as well: "[O]ur bodies are made up of more than gender and mere performance [77]." Performances are what these computer vision systems understand, and "gender" is a prominent structure by which they have been designed to see. The premise of computer vision elides the perspective that gender is subjective and internally held, and that gender performance is not always an indicator of gender. As one #agender user wrote: *"PRESENTATION ≠ IDENTITY."*

5.2 The Bias Propagated Through Third-Party Applications

As evidenced by the many differing computer vision services available, this technology is often designed in silo—their models, their data, and their labelling practices are generally proprietary blackboxes. However, even though our study focused on computer vision services, we cannot overlook their role as infrastructure and how the design of these services propagates into the applications that make use of them. In the previous section, we posit that the facial analysis and image labeling services reiterate archaic language about gender repackaged as neutral and technologically advanced. These services, designed to serve as infrastructure, have the potential to cascade into endless domains. In the hands of third-parties—where the neutral presentation of this worldview as a technological service might not be questioned—the notion of external gendered appearances as an indicator of a binary gender classification becomes calcified. It becomes embedded in numerous other infrastructures representing numerous other use cases. Here we can see many of the now familiar critiques of big data. However, in the context of these cloud-based services, there is a shift from data that is analyzed to *affordances* that are *used*.

As we have already discussed, labels present a potential alternative to rigid gender classifications and opportunity to embrace more diverse worldview. Yet, the diversity of the data returned by labeling services presents a challenge for third-party developers. In contrast to the data standard that gender classifiers offer to third-parties, the dynamic set of labels provided to developers provides a small, but not inconsequential, technical challenge. Labeling features only provide a

list of what was detected, not what wasn't, and it can be difficult to discern *why* specific labels manifest and others do not.

Our focus on how third-parties make use of these services is critical as that is where it is most likely that the choices embedded into cloud-based infrastructures will cause harm. Gender identification, particularly mandated through the state, has already been used to police trans identities (e.g., by barring trans individuals from accessing healthcare [58]). Social and physical harm could be perpetrated using computer vision technology to trans individuals, who already face high levels of harassment and violence [50, 112]. Binary gender classification in facial analysis could be used to intentionally obstruct access to social spaces (e.g. bathrooms ([11, 48])), restrict movement (e.g. the U.S. Transportation Security Administration (TSA) [26]), and even enact systemic and targeted violence if adopted by virulently anti-trans governments (e.g. ([49])).

Even if harming trans individuals is not an intentional outcome of a third-party system, interacting with tools that use the FA and image labeling infrastructure we studied could result in torque. It is not hard to imagine how the proliferation of large scale services like the ones we have studied could also scale experiences of misgendering documented by others [53, 74]. For example, the high rate of gender misclassification we observed, particularly for trans men, results in their identities as men being erased and twisted to fit into "female" classifications. Trans women, too, were frequently erased by misclassification, compounding the archaic and dangerous conflation of trans women as men in disguise [13, 114]. Given the rate at which trans individuals, even in our small dataset, discussed the emotional toil of dysphoria, designers should attend how insensitive FA classification could exacerbate the torque associated with both misgendering and dysphoria. The emotional harm—caused by misgendering and resulting in dysphoria—caused by gender misclassification can compound the torque already experienced by trans individuals on a daily basis.

Furthermore, for those who fell between these binary classifications altogether—existing only in residual categories that are not captured within the classification schema—the potential for torque is high. Binary classification forces non-binary users to conform to cisnormative expectations of gender performance. Non-binary genders were, metaphorically, molded to fit into two buckets of demographic gender (male versus female). Non-binary genders present a challenge for gender classification, but also highlight the challenges of designing human-centered systems built on computer vision infrastructure. In the eyes of these services (as they currently exist), human beings can only exist on a male/masculine versus female/feminine spectrum and that spectrum exists on a measurable, numerical probabilistic scale. As one #genderqueer person from our dataset posted in their caption: *"There is no right way to do gender, as long as you do it your way. Why settle for someone else's gender label when you can define your own?"* Yet, these services effectively assign "someone else's gender" to individuals. The opacity of these blackbox systems, and the limited understanding how gender classifications are being made, might also intensify torque. Individuals may not understand how their gender is being classified by the system, potentially resulting in increased self-doubt and negative affect about self-presentation.

It is difficult to predict how third-parties might use these commercial services, both in the present and in the future. Any number of use cases, intentionally harmful or not, could exist without the knowledge of the providers of this infrastructure. We have already seen instances of this in the alleged use of Microsoft Azure's by a Chinese company, SenseNets, to track Muslim minorities in Xinjiang [31]. But even beyond scenarios covered by popular press, it is likely that FA infrastructure is being used in countless smaller instances that may seem benign, but collectively reproduce a particular view of gender into our sociotechnical fabric. Political and social agendas, Bowker and Star remind us, are often first presented as purely technical interventions: "As layers of classification system become enfolded into a working infrastructure, the original political intervention becomes more and more firmly entrenched... It becomes taken for granted" [85]. With this in mind, it is

critical that designers, researchers, and policymakers think through designing the future of gender in facial analysis and image labeling services.

6 DESIGN AND POLICY CONSIDERATIONS

Understanding how gender is represented in facial analysis and image labeling infrastructure and how those representations might impact, in particular, trans individuals who come into contact with applications that use this infrastructure leads us to contemplate two key places to intervene: design and policy. In this section, we present implications for the design of computer vision models and datasets, as well as considerations for computer vision policies and standards.

6.1 Design of Facial Analysis and Image Labeling Services and its Applications

6.1.1 Use gender in classification carefully. The prevalence of gender classification across services may be an indicator that this is a feature that is important to and used by third-party clients. However, as the varied exclusion of race and ethnicity from services suggests, the creators of these services should consider why gender classification is being used in the first place. While this is perhaps obvious, we feel it is important to posit that designers carefully think what benefits gender brings to their system, and consider *abandoning* gender classification in facial analysis technology. Before embedding gender classification into a facial analysis service or incorporating gender into image labeling, it is important to consider what purpose gender is serving. Furthermore, it is important to consider how gender will be defined, and whether that perspective is unnecessarily exclusionary (e.g. binary). Binary gender should never be an unquestioned default. We propose that stakeholders involved in the development of facial analysis services and image datasets think through the potentially negative and harmful consequences their service might be used for—including emotional, social, physical, and systematic (state or governmental) harms.

6.1.2 Embrace gender ambiguity instead of the gender binary. When gender classification synthesizes gender performance into simplistic binary categories, the potential for gender fluidity and self-held gender identity is reduced. Labels like "person," "people," and "human" already provide inclusive information about the presence of human beings in a photograph. Rather than relying on static, binary gender in a face classification infrastructure, designers of applications should consider embracing, and demanding improvements, to feature-based labeling. Labels based on neutral performative markers (e.g. beard, makeup, dress) could replace gender classification in the facial analysis model, allowing third parties and individuals who come into contact with facial analysis applications to embrace their own interpretations of those features. This might actually be more precise, for the purposes of third-party applications. For example, performative markers like makeup would actually be more relevant to beauty product advertisers than gender classification, because they could then capture all genders who wear makeup. The multiplicity of labels does come with some technical overhead. Parsing and designing around a dynamic set of labels will always be more complex than simply checking for one of two values from a gender classifier. We acknowledge this, but also suggest that gender should not be simple.

6.1.3 Focus on contextualizing labeling. As we explicated in our discussion, computer vision services are currently unable to piece together contextual markers of identity. Rather than focusing on improving methods of gender classification, app designers could use labeling alongside other qualitative data, like the Instagram captions, to formulate more precise notions about user identity.

6.1.4 If gender must be used, consider the context of its application. If gender is something that is found to be useful to a system, designers should carefully consider the context the system will be used in. For example, while gender may be relevant to mitigating gender bias [118], consider what

kinds of bias are being privileged and what kinds of bias are being made invisible. Furthermore, consider the potential implications of gendered data being leaked, hacked, or misappropriated. For example, if attempting to mitigate bias against trans individuals, consider whether attempting to explicitly embed trans gender recognition into a model could do more harm than good. This presents a tension. On one hand, gender classification could be used for the benefit of mitigating gender bias—through recognizing performative markers of underrepresented genders. On the other hand, the same system could be adopted in a way that undermines the benefits and results in harm. Service providers should consider how to provide gender classification functionality to third-party developers in a way that enables more scrutiny and oversight.

6.2 Design of Image Datasets

Some have appealed for gender inclusive datasets, including images of trans people with diverse genders (e.g. [65]; others are concerned with the implications of training facial analysis services to recognize trans identities (e.g. [51, 94]). We urge designers to consider the risks of training computer vision to identify trans individuals in an attempt to be more gender inclusive. Our current work highlights the challenges involved in creating an inclusive dataset—and, in fact, argues that a truly, universally inclusive dataset is not possible. With that in mind, we recommend three approaches to consider towards developing more inclusive image training datasets. In all cases, designers should make explicit exactly what the data is being used for, and ensure not to sell that data to other parties who might use it for harm.

6.2.1 Use self-identified gender in datasets. When gender classification is appropriate, it is important for it to be accurate. Like we found in our analysis, the same image might be classified differently across computer vision services. Primarily, gender labeling practices currently require labeling to be done based off of subjective interpretation of external appearance. For something as complex and personal as gender, relying on datasets where human labelers have inferred gender leads to inaccuracies. However, the caveat to ignoring gender in datasets is potentially reifying gender bias (e.g. ([24, 60])). If the computer vision application requires gender, creating a dataset of self-identified gender could mitigate some bias inherent in subjective labeling. To do this, designers should seek explicit consent from individuals to use their images and label data through a continuous informed consent process. However, given that computer vision is limited to what can be seen, designers might find it necessary to build systems that rely on more than just computer vision and make use of other forms of data. Given that computer vision is limited to what can be seen, designers might find it necessary to build services that rely on more than just computer vision and make use of other forms of data.

6.2.2 Consider the tensions of gender classification annotations in datasets. Whether gender is necessary to include in the infrastructure of a computer vision model should determine how gender should be built into data labeling practices at all. In use cases where the classifiers' purpose is to mitigate bias, gender labeled data would be necessary for the model to function. For example, a classifier built for the purposes of trying to improve gender parity in hiring women. Trans women are also women, so should their images be labeled as "woman"? However, trans individuals are also underrepresented in hiring [50] and should be accounted for. This would require explicit trans labeling, which could be an issue of consent (in which trans women do not wish to be outed as trans) and open the doors for potential misuse of the system (intentionally not hiring trans women).

6.2.3 Focus on including a diverse set of performative gender labels. One method to consider when building a diverse and inclusive dataset is constructing a heuristic for diverse gender markers across a range of skin tones. In doing this, designers should consider creating datasets that allow multiple

performative values to exist. For example, working to develop a range of gender markers that could overlap (e.g. beards, long hair, makeup, clothing style) using participatory design methods with gender diverse individuals (including cisgender and trans individuals).

6.3 Policies and Regulations

There has already been an increasing call for policy regulation for how facial analysis technologies are built and used [69, 111]; some governments have already moved towards enacting such policies (e.g., [84, 96, 100]). Considering long-standing state policies around gender identification are recently changing (with the advent of legally permissible 'X' gender markers (e.g., [29, 35, 102])), current gender representations in commercial computer vision services are already obsolete in the United States, where these companies are based. We recommend future-looking policy considerations for gender classification in facial analysis and image labeling.

6.3.1 Create policies for inclusive standards for how gender is used in computer vision systems. As we found in section 3.1, gender is used in some services but not all. When it is present, it is not consistent across all services. We recommend that relevant stakeholders—including designers, policymakers, engineers, and researchers of diverse genders—work to establish principled guidelines towards gender inclusivity for computer vision infrastructures. Not only would such a policy promote equity in gender representation, it would make bias auditing and harm mitigation for facial analysis and image labeling services easier. It would also benefit third-party developers who need to move between services. As the concept of gender is shifting, both socially and legally, we'd also recommend reassessing policies and guidelines regulating how gender is used by computer vision systems regularly.

6.3.2 Establish policies to hold companies accountable for how services are used. Currently service providers are often not held accountable for how their services are used, but we are starting to see a shift in public expectations (e.g. [83]). However, the current architecture of these services may limit the services ability to understand the ultimate purpose or use of the third-party system (e.g. ([31])). We acknowledge the significant challenges here, however, it is important to develop policies that establish layers of accountability and transparency for how FA and image labeling services are used by third-party applications to ensure that identity classification in computer vision services are not used to perpetrate harm.

6.3.3 Treat trans identity as a protected class. Much like how the Fair Housing Act extends its anti-discrimination policies to online advertisers [73], policymakers can consider how to expand legally "protected classes" to encompass facial analysis technologies. Establishing policies for how biometric data and face and body images are collected and used may be the most effective way of mitigating harm to trans people—and also people of marginalized races, ethnicities, and sexualities. Polices that prevent discriminatory and non-consensual gender representations could prevent gender misrepresentation from being incorporated into FA systems in both the data and infrastructure by regulating the use of gender as a category in algorithmic systems. For example, by banning the use of gender from FA-powered advertising and marketing.

7 LIMITATIONS AND FUTURE WORK

At many points in this study, we reached the limitations of our methods—we cannot see inside these black boxes. However, an inside perspective is crucial for future scholarship. Conducting research on how gender is embedded into these services throughout their development pipeline, particularly by talking with designers and practitioners who are developing current systems, is necessary for a deeper understanding of why, when, and how gender is conceptualized for computer

visions services. Future work should focus on uncovering the motivations and rationale behind the development of gender classification in commercial facial analysis and labeling services, and the points of translation through which gender moves from a complex social concept to a data point amenable to computation.

Furthermore, while we sought a diverse representation of binary and non-binary genders (including genderless "genders," i.e., agender), there is boundless opportunity to include other genders in computer vision research. We also briefly discussed in section 4.2.1 that assessing FA services required assumptions to be made about gender identity and pronouns. We recognize this as a limitation, still rooted in binary conceptions of gender that assume trans men use he/him and would be situated in "male" classification categories. Also, while we did not evaluate the impact of skin tone or ethnicity, knowing that skin tone impacts classification performance of gender classification in facial analysis software [19], future work would benefit from analyzing gender diversity alongside skin tone and ethnicity. Certainly, more diverse genders and skin tones should be included in imagining ethical solutions to categorization schemas. Future work might also explore different measurements of accuracy and performance on diverse gender datasets. While sufficient for the purposes of this study, we also recognize that a dataset of 2450 images, sub-divided into seven 350 gender datasets, is rather small in the world of machine learning. We believe that larger datasets with more diverse genders and skin tones would provide new and interesting insights to this research domain.

8 CONCLUSION

Current research on gender classification in computer vision services, specifically with facial analysis technologies, has unearthed crucial issues of racial bias [19] and trans representations in current automatic gender recognition approaches [46, 57]. Our study builds on the inroads these researchers have already paved by providing empirical evidence to support their findings about how gender is handled in gender recognition systems. We directly examined how (1) commercial computer vision services classify and label images of different genders, including non-binary genders, as well as how (2) labeling constructs a cultural reality of gender within computer vision infrastructure.

To do this, we constructed a dataset of photos including diverse genders to demonstrate how these services see—and are unable to see—both binary and non-binary genders. Through a systems analysis of these services, quantitative evaluation of gender classification, and a qualitative analysis of images labeling, we provide new insights into how computer vision services operationalize gender. We found that binary gender classification provided by computer vision services performed worse on binary trans images than cis ones, and were unable to correctly classify non-binary genders. While image labeling differed by providing labels that allowed for gender neutrality (e.g., "person") or multiplicity (e.g., "man" and "woman"), they still made use of a binary notion of gender performance.

We discussed how different perspectives are encoded in cloud-based infrastructure that propagate into software developed by third-parties, potentially resulting in harm to the individuals who interact with technology that uses this infrastructure. Throughout we have highlighted the importance of considering how gender classification becomes mediated across technological layers—from infrastructure, to third-party developers, to end users. We conclude with recommendations for designing infrastructure and datasets, and outline implications for policy that would improve inclusivity and mitigate potential harm.

ACKNOWLEDGMENTS

We would like thank the reviewers for their time and dedication towards improving this paper. We would also like to thank the CU Boulder HCC community for their valuable feedback in the process of writing this paper — in particular, Michael Paul and Casey Fiesler, whose feedback helped strengthen our methodological considerations. Finally, thank you to Aaron Jiang for helping us whenever we struggled with formatting tables in Overleaf. This work was supported in part by the National Science Foundation #1042678.

REFERENCES

[1] 2019. Amazon Rekognition – Video and Image - AWS. https://aws.amazon.com/rekognition/
[2] 2019. Clarifai. https://clarifai.com/
[3] 2019. Face API - Facial Recognition Software | Microsoft Azure. https://azure.microsoft.com/en-us/services/cognitive-services/face/
[4] 2019. Face++ Cognitive Services - Leading Facial Recognition Technology. https://www.faceplusplus.com/
[5] 2019. Vision API - Image Content Analysis | Cloud Vision API | Google Cloud. https://cloud.google.com/vision/
[6] 2019. Watson Visual Recognition. https://www.ibm.com/watson/services/visual-recognition/
[7] Yaman Akbulut, Abdulkadir Sengur, and Sami Ekici. 2017. Gender recognition from face images with deep learning. In *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, 1–4. https://doi.org/10.1109/idap.2017.8090181
[8] Tawfiq Ammari, Sarita Schoenebeck, and Silvia Lindtner. 2017. The Crafting of DIY Fatherhood. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. ACM Press, New York, New York, USA, 1109–1122. https://doi.org/10.1145/2998181.2998270
[9] Ankan Bansal, Anirudh Nanduri, Carlos D. Castillo, Rajeev Ranjan, and Rama Chellappa. 2018. UMDFaces: An annotated face dataset for training deep networks. In *IEEE International Joint Conference on Biometrics, IJCB 2017*, Vol. 2018-Janua. IEEE, 464–473. https://doi.org/10.1109/BTAS.2017.8272731
[10] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*. ACM Press, New York, New York, USA, 1301. https://doi.org/10.1145/1753326.1753521
[11] Kyla Bender-Baird. 2015. Peeing under surveillance: bathrooms, gender policing, and hate violence. *Gender, Place & Culture* 23, 7 (jul 2015), 983–988. https://doi.org/10.1080/0966369x.2015.1073699
[12] Sebastian Benthall and Bruce D. Haynes. 2019. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, New York, New York, USA, 289–298. https://doi.org/10.1145/3287560.3287575 arXiv:1811.11668
[13] Talia Mae Bettcher. 2007. Evil Deceivers and Make-Believers: On Transphobic Violence and the Politics of Illusion. *Hypatia* 22, 3 (aug 2007), 43–65. https://doi.org/10.1111/j.1527-2001.2007.tb01090.x
[14] Rena Bivens. 2014. The Gender Binary Will Not Be Deprogrammed: Facebook's Antagonistic Relationship to Gender. *SSRN Electronic Journal* (dec 2014). https://doi.org/10.2139/ssrn.2431443
[15] Rena Bivens and Oliver L. Haimson. 2016. Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society* 2, 4 (nov 2016), 1–12. https://doi.org/10.1177/2056305116672486
[16] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob ACM Reference format. *Proc. ACM Hum.-Comput. Interact. Proc. ACM Hum.-Comput. Interact. Article Proc. ACM Hum.-Comput. Interact* 1, 2 (2017), 1–19. https://doi.org/10.1145/3134659
[17] Brad Smith. 2018. Facial recognition: It's time for action. https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/
[18] Jed R. Brubaker and Gillian R. Hayes. 2011. SELECT * FROM USER: infrastructure and socio-technical representation. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*. ACM Press, New York, New York, USA, 369. https://doi.org/10.1145/1958824.1958881
[19] Joy Buolamwini and Timnit Gebru. 2018. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* *. Technical Report. 1–15 pages. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
[20] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. , 202–214 pages. http://proceedings.mlr.press/v81/burke18a.html
[21] Judith Butler. 1988. Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory. *Theatre Journal* 40, 4 (dec 1988), 519. https://doi.org/10.2307/3207893

[22] Ting Chen, Wei Li Han, Hai Dong Wang, Yi Xun Zhou, Bin Xu, and Bin Yu Zang. 2007. Content recommendation system based on private dynamic user profile. In *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007*, Vol. 4. IEEE, 2112–2118. https://doi.org/10.1109/ICMLC.2007.4370493

[23] John Cheney-Lippold. 2011. A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control. *Theory, Culture & Society* 28, 6 (nov 2011), 164–181. https://doi.org/10.1177/0263276411424420

[24] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. (jul 2018). https://doi.org/10.1063/1.3627170 arXiv:1808.00023

[25] Critical Art Ensemble. 1998. *Flesh Machine.* http://critical-art.net/flesh-machine-1997-98/

[26] Paisley Currah and Tara Mulqueen. 2011. Securitizing Gender: Identity, Biometrics, and Transgender Bodies at the Airport. *Social Research* 78, 2 (2011), 557–582. https://doi.org/10.1353/sor.2011.0030

[27] Ya E. Dai, Hong Wu Ye, and Song Jie Gong. 2009. Personalized recommendation algorithm using user demography information. In *Proceedings - 2009 2nd International Workshop on Knowledge Discovery and Data Mining, WKKD 2009.* IEEE, 100–103. https://doi.org/10.1109/WKDD.2009.156

[28] Avery Dame. 2016. Making a name for yourself: tagging as transgender ontological practice on Tumblr. *Critical Studies in Media Communication* 33, 1 (jan 2016), 23–37. https://doi.org/10.1080/15295036.2015.1130846

[29] Scott Dance. 2019. Maryland set to add 'X' gender designation to driver's licenses under bill by General Assembly. https://www.baltimoresun.com/news/maryland/politics/bs-md-drivers-licenses-20190313-story.html

[30] Heath Fogg Davis. 2017. *Beyond Trans: Does Gender Matter?* https://books.google.com/books?id=uHA4DQAAQBAJ{&}source=gbs{_}navlinks{_}s

[31] Zak Doffman. 2019. Is Microsoft AI Helping To Deliver China's 'Shameful' Xinjiang Surveillance State? https://www.forbes.com/sites/zakdoffman/2019/03/15/microsoft-denies-new-links-to-chinas-surveillance-state-but-its-complicated/{#}4cb624f73061

[32] Grant Duwe and Ki Deuk Kim. 2017. Out With the Old and in With the New? An Empirical Comparison of Supervised Learning Algorithms to Predict Recidivism. *Criminal Justice Policy Review* 28, 6 (jul 2017), 570–600. https://doi.org/10.1177/0887403415604899

[33] Brianna Dym, Jed Brubaker, and Casey Fielser. 2018. "theyre all trans sharon": Authoring Gender in Video Game Fan Fiction. *Game Studies* 3, 18 (2018). http://gamestudies.org/1803/articles/brubaker{_}dym{_}fieslerhttp://gamestudies.org/1701/articles/anderson

[34] W. Keith Edwards, Mark W. Newman, and Erika Shehan Poole. 2010. The infrastructure problem in HCI. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10.* ACM Press, New York, New York, USA, 423. https://doi.org/10.1145/1753326.1753390

[35] Elise Schmelzer. 2018. Colorado to allow use of X as sex identifier on driver's licenses starting this month.

[36] Erik Carter. 2019. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

[37] Erik H. (Erik Homburger) Erikson and Joan M. (Joan Mowat) Erikson. 1982. The life cycle completed. (1982), 134. https://www.worldcat.org/title/life-cycle-completed/oclc/916049006

[38] Melanie Feinberg, Daniel Carter, and Julia Bullard. 2014. A Story Without End : Writing the Residual into Descriptive Infrastructure. *DIS '14 Proceedings of the Designing Interactive Systems Conference* (2014), 385–394. https://doi.org/10.1145/2598510.2598553

[39] Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media and Society* 4, 1 (jan 2018), 205630511876336. https://doi.org/10.1177/2056305118763366

[40] Brian Joseph Gilley. 2016. Imagining Transgender: An Ethnography of a Category . David Valentine. In *Journal of Anthropological Research.* Vol. 65. Duke University Press, Chapter Imagining, 516–517. https://doi.org/10.1086/jar.65.3.25608249

[41] Erving. Goffman. 1956. The Presentation of Self in Everyday Life. *The Production of Reality: Essays and Readings on Social Interaction* (1956), 262. https://books.google.com/books/about/The{_}Presentation{_}of{_}Self{_}in{_}Everyday{_}Lif.html?id=Sdt-cDkV8pQC

[42] Patrick Grother, Mei Ngan, Kayee Hanaoka, and Wilbur L Ross. 2018. Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification. *NIST Interagency/Internal Report (NISTIR)* (nov 2018). https://doi.org/10.6028/NIST.IR.8238

[43] Oliver L. Haimson, Jed R. Brubaker, Lynn Dombrowski, and Gillian R. Hayes. 2015. Disclosure, Stress, and Support During Gender Transition on Facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15).* ACM Press, New York, New York, USA, 1176–1190. https://doi.org/10.1145/2675133.2675152

[44] Oliver L. Haimson and Anna Lauren Hoffmann. 2016. Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities. *First Monday* 21, 6 (jun 2016). https://doi.org/10.5210/fm.v21i6.6791

[45] Jack Halberstam. 1998. *Female Masculinity.* https://books.google.com/books/about/Female{_}Masculinity.html?id=5BqOswEACAAJ{&}source=kp{_}book{_}description

[46] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Automatic Gender Recognition Systems. In *2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*.

[47] D. Fox Harrell. 2009. Computational and cognitive infrastructures of stigma. In *Proceeding of the seventh ACM conference on Creativity and cognition - C&C '09*. ACM Press, New York, New York, USA, 49. https://doi.org/10.1145/1640233.1640244

[48] Jody L Herman. 2013. Gendered restrooms and minority stress: The public regulation of gender and its impact on transgender people's lives. *Journal of Public Management and Social Policy* (2013), 65–80. https://williamsinstitute.law.ucla.edu/wp-content/uploads/Herman-Gendered-Restrooms-and-Minority-Stress-June-2013.pdf

[49] Marie Hicks. 2019. Hacking the Cis-tem: Transgender Citizens and the Early Digital State. *IEEE Annals of the History of Computing* 41, 1 (jan 2019), 1–1. https://doi.org/10.1109/mahc.2019.2897667

[50] Sandy E. James, Jody L. Herman, Susan Rankin, Mara Keisling, Lisa Mottet, and Ma'ayan Anafi. 2016. *The Report of the 2015 U.S. Transgender Survey*. Technical Report. National Center for Transgender Equality. 298 pages. http://www.transequality.org/sites/default/files/docs/usts/USTSFullReport-FINAL1.6.17.pdf

[51] James Vincent. 2017. Transgender YouTubers had their videos grabbed to train facial recognition software. https://www.theverge.com/2017/8/22/16180080/transgender-youtubers-ai-facial-recognition-dataset

[52] James Vincent. 2019. AI researchers tell Amazon to stop selling 'flawed' facial recognition to the police. https://www.theverge.com/2019/4/3/18291995/amazon-facial-recognition-technology-rekognition-police-ai-researchers-ban-flawed?fbclid=IwAR2BE5ObzjkVeq5W6hlPVobNsyEYCfAvIYZV6Jq-0I0HlErQkxniHpHoDIc

[53] Stephanie Julia Kapusta. 2016. Misgendering and Its Moral Contestability. *Hypatia* 31, 3 (aug 2016), 502–519. https://doi.org/10.1111/hypa.12259

[54] Soon-Gyo Jung, Jisun An, Haewoon Kwak, Joni Salminen, and Bernard Jim Jansen. 2018. Assessing the Accuracy of Four Popular Face Recognition Tools for Inferring Gender, Age, and Race. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*. 624–627. https://www.cnet.com/news/google-apologizes-for-algorithm-https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17839

[55] Yannis Kalantidis, Munmun De Choudhury, Jessica A. Pater, Stevie Chancellor, and David A. Shamma. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM Press, New York, New York, USA, 3213–3226. https://doi.org/10.1145/3025453.3025985

[56] Jakko Kemper and Daan Kolkman. 2018. Transparent to whom? No algorithmic accountability without a critical audience. , 16 pages. https://doi.org/10.1080/1369118X.2018.1477967

[57] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (nov 2018), 1–22. https://doi.org/10.1145/3274357

[58] Liza Khan. 2011. Transgender Health at the Crossroads: Legal Norms, Insurance Markets, and the Threat of Healthcare Reform. *Yale Journal of Health Policy, Law & Ethics* 11, c (2011), 375–418. https://heinonline.org/HOL/Page?handle=hein.journals/yjhple11{&}id=381{&}collection=journals{&}index=

[59] Sajid Ali Khan, Maqsood Ahmad, Muhammad Nazir, and Naveed Riaz. 2013. A comparative analysis of gender classification techniques. *International Journal of Bio-Science and Bio-Technology* 5, 4 (2013), 223–243. https://doi.org/10.5829/idosi.mejsr.2014.20.01.11434

[60] Anja Lambrecht and Catherine E. Tucker. 2016. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. (2016). https://doi.org/10.2139/ssrn.2852260

[61] Alex Leavitt. 2015. "This is a Throwaway Account": Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. ACM Press, New York, New York, USA, 317–327. https://doi.org/10.1145/2675133.2675175

[62] Hongjun Li and Ching Y. Suen. 2016. Robust face recognition based on dynamic rank representation. *Pattern Recognition* 60, C (dec 2016), 13–24. https://doi.org/10.1016/j.patcog.2016.05.014

[63] Stan Z. Li and Anil K. Jain. 2011. *Handbook of Face Recognition*. https://doi.org/10.1007/978-0-85729-932-1

[64] Yang Li, Suhang Wang, Jiliang Tang, Quan Pan, and Tao Yang. 2017. Price Recommendation on Vacation Rental Websites. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. 399–407. https://doi.org/10.1137/1.9781611974973.45

[65] Lindsay Schrupp. 2019. Why We Created a Gender-Inclusive Stock Photo Library. https://broadly.vice.com/en{_}us/article/qvyq8p/transgender-non-binary-stock-photos-gender-spectrum-collection

[66] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2015 Inter. 3730–3738. https://doi.org/10.1109/ICCV.2015.425 arXiv:1411.7766

[67] Xiaoguang Lu and Anil K Jain. 2004. Ethnicity Identification from Face Images. *Proceedings of SPIE* 5404 (2004), 114–123. https://doi.org/10.1117/12.542847

[68] Gayathri Mahalingam and Karl Ricanek. 2013. Is the eye region more reliable than the face? A preliminary study of face-based recognition on a transgender dataset. In *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS 2013)*. IEEE, 1–7. https://doi.org/10.1109/BTAS.2013.6712710

[69] Makena Kelly. 2019. Pressure mounts on Google, Microsoft, and Amazon's facial recognition tech. https://www.theverge.com/2019/1/15/18183789/google-amazon-microsoft-pressure-facial-recognition-jedi-pentagon-defense-government

[70] James E Marcia. 1966. *Development and Validation of Ego-Identity Status*. Ph.D. Dissertation. https://pdfs.semanticscholar.org/f145/f3fbada1eb7a01052255f586094301669287.pdf

[71] Annette Markham. 2012. Fabrication as ethical practice: Qualitative inquiry in ambiguous Internet contexts. *Information Communication and Society* 15, 3 (apr 2012), 334–353. https://doi.org/10.1080/1369118X.2011.641993

[72] Matthew Gault. 2019. Facial Recognition Software Regularly Misgenders Trans People. https://motherboard.vice.com/en{_}us/article/7xnwed/facial-recognition-software-regularly-misgenders-trans-people

[73] John D. McKinnon and Jeff Horwitz. 2019. HUD Action Against Facebook Signals Trouble for Other Platforms. https://www.wsj.com/articles/u-s-charges-facebook-with-violating-fair-housing-laws-11553775078

[74] Kevin A. McLemore. 2015. Experiences with Misgendering: Identity Misclassification of Transgender Spectrum Individuals. *Self and Identity* 14, 1 (jan 2015), 51–74. https://doi.org/10.1080/15298868.2014.950691

[75] Cade Metz. 2019. Is Ethical A.I. Even Possible? https://www.nytimes.com/2019/03/01/business/ethics-artificial-intelligence.htmlhttps://www.nytimes.com/2019/03/01/business/ethics-artificial-intelligence.html?utm{_}source=Benedict{%}27s+newsletter{&}utm{_}campaign=a333b6b622-Benedict{%}27s+Newsletter{_}COPY{_}01{&}utm{_}

[76] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R Varshney. 2018. Understanding Unequal Gender Classification Accuracy from Face Images. (2018). arXiv:1812.00099 https://github.com/ox-vgg/vgg{_}face2http://arxiv.org/abs/1812.00099

[77] Viviane K. Namaste. 2006. Invisible Lives: The Erasure of Transsexual and Transgendered People. *Contemporary Sociology* 31, 3 (2006), 264. https://doi.org/10.2307/3089651

[78] Natalia Drozdiak. 2019. Microsoft Seeks to Restrict Abuse of its Facial Recognition AI. https://www.bloomberg.com/news/articles/2019-01-23/microsoft-seeks-to-restrict-abuse-of-its-facial-recognition-ai

[79] A.J. Neuman Wipfler. 2016. Identity Crisis: the Limitations of Expanding Government Recognition of Gender Identity and the Possibility of Genderless Identity Documents. *Harvard Journal of Law and Gender* 39, 2 (2016), 401–464. https://heinonline.org/HOL/Page?handle=hein.journals/hwlj39{&}id=505{&}collection=journals{&}index=

[80] Choon Boon Ng, Yong Haur Tay, and Bok Min Goi. 2015. A review of facial gender recognition. *Pattern Analysis and Applications* 18, 4 (nov 2015), 739–755. https://doi.org/10.1007/s10044-015-0499-6

[81] Mei Ngan and Patrick Grother. 2015. *Face Recognition Vendor Test (FRVT) - Performance of Automated Gender Classification Algorithms*. Technical Report. https://doi.org/10.6028/NIST.IR.8052

[82] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, New York, New York, USA, 89–89. https://doi.org/10.1145/3287560.3287593

[83] Sara Ashley O'Brien. [n. d.]. What is Amazon's responsibility over its facial recognition tech? https://money.cnn.com/2018/07/26/technology/amazon-facial-recognition/index.html

[84] Joseph O'Sullivan. 2019. Washington Senate approves consumer-privacy bill to place restrictions on facial recognition. https://www.seattletimes.com/seattle-news/politics/senate-passes-bill-to-create-a-european-style-consumer-data-privacy-law-in-washington/

[85] J. Marc Overhage and Jeffery G. Suico. 2013. Sorting Things Out: Classification and Its Consequences. *Annals of Internal Medicine* 135, 10 (2013), 934. https://doi.org/10.7326/0003-4819-135-10-200111200-00030 arXiv:arXiv:1011.1669v3

[86] P Jonathon Phillips, Abhijit Narvekar, Alice J. O'Toole, Fang Jiang, and Julianne Ayyad. 2011. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception* 8, 2 (2011), 1–11. https://doi.org/10.1145/1870076.1870082

[87] Mark Poster. 2013. *The Mode of Information: Poststructuralism and Social Context*. Polity Press. https://doi.org/10.5860/crl_52_03_300

[88] Rachel Metz. 2019. Amazon shareholders want it to stop selling facial-recognition tech to the government. https://www.cnn.com/2019/01/17/tech/amazon-shareholders-facial-recognition/index.html

[89] Inioluwa Deborah Raji and Joy Buolamwini. 2019. *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*. Technical Report. 7 pages. www.aaai.org

[90] Arnaud Ramey and Miguel A. Salichs. 2014. Morphological Gender Recognition by a Social Robot and Privacy Concerns. *Proceedings of the 2014 ACM/IEEE International conference on Human-Robot Interaction (HRI '14)* (2014), 272–273. https://doi.org/10.1145/2559636.2563714

[91] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6, 1 (dec 2017), 15. https://doi.org/10.1140/epjds/s13688-017-0110-z arXiv:1608.03282

[92] Jennifer A Rode. 2011. A theoretical agenda for feminist HCI. *Interacting with Computers* 23, 5 (2011), 393–400. https://doi.org/10.1016/j.intcom.2011.04.005

[93] Pau Rodríguez, Guillem Cucurull, Josep M. Gonfaus, F. Xavier Roca, and Jordi Gonzàlez. 2017. Age and gender recognition in the wild with deep attention. *Pattern Recognition* 72 (dec 2017), 563–571. https://doi.org/10.1016/J.PATCOG.2017.06.028

[94] Janus Rose. 2019. I'm a trans woman – here's why algorithms scare me | Dazed. https://www.dazeddigital.com/science-tech/article/43211/1/trans-algorithm-machine-learning-bias-discrimination-chelsea-manning-edit

[95] Gayle Rubin. 2013. Literary theory: an anthology. In *Choice Reviews Online*, Julie Rivkin and Michael Ryan (Eds.). Vol. 35. Chapter The Traffi, 35–5478–35–5478. https://doi.org/10.5860/choice.35-5478

[96] Russell Brandom. 2019. Crucial biometric privacy law survives Illinois court fight. https://www.theverge.com/2019/1/26/18197567/six-flags-illinois-biometric-information-privacy-act-facial-recognition

[97] Manisha M. Sawant and Kishor M. Bhurchandi. 2018. Age invariant face recognition: a survey on facial aging databases, techniques and effect of aging. *Artificial Intelligence Review* (oct 2018), 1–28. https://doi.org/10.1007/s10462-018-9661-z

[98] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proceedings of the ACM on Human-Computer Interaction* 2 (2018), 29.

[99] Ari Schlesinger, Christina A. Masden, Rebecca E. Grinter, Eshwar Chandrasekharan, W. Keith Edwards, and Amy S. Bruckman. 2017. Situated Anonymity: Impacts of Anonymity, Ephemerality, and Hyper-Locality on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM Press, New York, New York, USA, 6912–6924. https://doi.org/10.1145/3025453.3025682

[100] Sidney Fussell. 2019. San Francisco Wants to Ban Government Face Recognition. https://www.theatlantic.com/technology/archive/2019/02/san-francisco-proposes-ban-government-face-recognition/581923/

[101] Jacob Snow. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots.

[102] Brooke Sopelsa. 2018. Gender 'X': New York City to add third gender option to birth certificates. https://www.nbcnews.com/feature/nbc-out/gender-x-new-york-city-add-third-gender-option-birth-n909021

[103] Susan Stryker. 2006. (De)subjugated Knowledges: An Introduction to Transgender Studies. In *The Transgender Studies Reader*. Routledge, 1–15. https://doi.org/10.4324/9780203955055-7

[104] G Subbalakshmi. 2011. Decision Support in Heart Disease Prediction System using Naive Bayes. *Indian Journal of Computer …* (2011). http://www.ijcse.com/docs/IJCSE11-02-02-56.pdf

[105] Tom Simonite. 2019. Microsoft Wants Rules for Facial Recognition—Just Not These. https://www.wired.com/story/microsoft-wants-rules-facial-recognition-just-not-these/

[106] E. B. Towle. 2005. ROMANCING THE TRANSGENDER NATIVE: Rethinking the Use of the "Third Gender" Concept. *GLQ: A Journal of Lesbian and Gay Studies* 8, 4 (jan 2005), 469–497. https://doi.org/10.1215/10642684-8-4-469

[107] Jacques Veron, Samuel H. Preston, Patrick. Heuveline, and Michel Guillot. 2006. Demography: Measuring and Modeling Population Processes. *Population (French Edition)* 57, 3 (2006), 591. https://doi.org/10.2307/1535065

[108] Shui Hua Wang, Preetha Phillips, Zheng Chao Dong, and Yu Dong Zhang. 2018. Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm. *Neurocomputing* 272 (jan 2018), 668–676. https://doi.org/10.1016/j.neucom.2017.08.015

[109] Yilin Wang, Neil O'Hare, Baoxin Li, Yali Wan, Jiliang Tang, Jundong Li, Clayton Mellina, and Yi Chang. 2017. Understanding and Discovering Deliberate Self-harm Content in Social Media. In *Proceedings of the 26th International Conference on World Wide Web - WWW '17*. ACM Press, New York, New York, USA, 93–102. https://doi.org/10.1145/3038912.3052555

[110] Laurel Westbrook and Kristen Schilt. 2013. Doing Gender, Determining Gender. *Gender & Society* 28, 1 (feb 2013), 32–57. https://doi.org/10.1177/0891243213503203

[111] Will Knight. 2018. Facial recognition has to be regulated to protect the public, says AI report. https://www.technologyreview.com/s/612552/facial-recognition-has-to-be-regulated-to-protect-the-public-says-ai-report/

[112] O. Wilson. 2013. Violence and Mental Health in the Transgender Community. December (2013). https://search.proquest.com/docview/1647175438?pq-origsite=gscholar

[113] Stelios C. Wilson, Shane D. Morrison, Lavinia Anzai, Jonathan P. Massie, Grace Poudrier, Catherine C. Motosko, and Alexes Hazen. 2018. Masculinizing Top Surgery: A Systematic Review of Techniques and Outcomes. , 679–683 pages. https://doi.org/10.1097/SAP.0000000000001354

[114] Aimee Wodda and Vanessa Panfil. 2015. "Don't talk to me about deception": The necessary erosion of the trans* panic defense. *Albany Law Review* 78, 3 (2015), 927–971. https://doi.org/10.1017/CBO9781107415324.004 arXiv:arXiv:1011.1669v3

[115] Steve Woolgar and Lucy Alice. Suchman. 1989. Plans and Situated Actions: The Problem of Human Machine Communication. *Contemporary Sociology* 18, 3 (1989), 414. https://doi.org/10.2307/2073874

[116] Billy Yapriady and Alexandra L. Uitdenbogerd. 2010. Combining Demographic Data with Collaborative Filtering for Automatic Music Recommendation. Springer, Berlin, Heidelberg, 201–207. https://doi.org/10.1007/11554028_29

[117] Ni Zhuang, Yan Yan, Si Chen, Hanzi Wang, and Chunhua Shen. 2018. Multi-label learning based deep transfer neural network for facial attribute classification. *Pattern Recognition* 80 (2018), 225–240. https://doi.org/10.1016/j.patcog.2018.03.018 arXiv:arXiv:1805.01282v1

[118] Indre Zliobaite and Bart Custers. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law* 24, 2 (jun 2016), 183–201. https://doi.org/10.1007/s10506-016-9182-5

A APPENDIX: TERMS OF SERVICE

Below is the information provided by the services on their websites regarding data storage and use of customer image data for model training. Pieces of these statements may be located on different webpages or within different terms of services on the company websites. We provide information for the five services we used during our in-depth study. Information from Face++ is provided as a contrasting example of terms of service that resulted in us excluded a service from our study.

A.0.1 Amazon Rekognition: "Amazon Rekognition may store and use image and video inputs processed by the service solely to provide and maintain the service and, unless you opt out as provided below, to improve and develop the quality of Amazon Rekognition and other Amazon machine-learning/artificial-intelligence technologies ... You can request deletion of image and video inputs associated with your account by contacting AWS Support. Deleting image and video inputs may degrade your Amazon Rekognition experience."

A.0.2 Clarifai. "Account termination may result in destruction of any User Content associated with your account, so keep that in mind before you decide to terminate your account ... When media is submitted to Clarifai for analysis, the media bytes (input) and tags (input) are sent to Clarifai's Kubernetes cluster of Amazon Web Services EC2 instances and discarded once the prediction is complete. If the media bytes were submitted via URL, the image has not been and will not be stored by Clarifai. Clarifai stores the concepts identified in the image bytes (outputs) in Citus Cloud Postgres Database and related account information in AWS Relational Database Service. This is required to present the results of the request to the customer."

A.0.3 Google Vision. "Google does not use any of your content (such as images and labels) for any purpose except to provide you with the Cloud Vision API service ... When you send an image to Cloud Vision API, we must store that image for a short period of time in order to perform the analysis and return the results to you. The stored image is typically deleted in a few hours. Google also temporarily logs some metadata about your Vision API requests (such as the time the request was received and the size of the request) to improve our service and combat abuse ... Google will enable Customer to delete Customer Data during the Term in a manner consistent with the functionality of the Services. If Customer uses the Services to delete any Customer Data during the Term and that Customer Data cannot be recovered by Customer, this use will constitute an instruction to Google to delete the relevant Customer Data from Google's systems in accordance with applicable law."

A.0.4 IBM Watson. "By default, all Watson services log requests and their results. Logging is done only to improve the services for future users. The logged data is not shared or made public. To prevent IBM usage of your data for an API request, set the X-Watson-Learning-Opt-Out header parameter to

true. You must set the header on each request that you do not want IBM to access for general service improvements."

A.0.5 Microsoft Azure. "Under the new terms, Cognitive Services customers own, and can manage and delete their customer data. With this change, many Cognitive Services are now aligned with the same terms that apply to other Azure services ... You can also make choices about the collection and use of your data by Microsoft. You can control your personal data that Microsoft has obtained, and exercise your data protection rights, by contacting Microsoft or using various tools we provide."

A.0.6 Face++ (Contrasting Example). "You hereby grant to us all rights to use and incorporate any contents You provide to us through Face++, apart from legal negotiations, confidential material or user login details, in the API Documents, or any other Face++ product or service for improving the Face++ APIs or any application that we have now or may develop in the future without compensation to You and without further recourse by You. Please be assured that, in respect of any photo that may have uploaded by You on Face++ or through the Face++ APIs, we will collect and save the original photo securely for our research purpose, not share with 3rd party."