To "See" is to Stereotype

Image Tagging Algorithms, Gender Recognition, and the Accuracy – Fairness Trade-off

PINAR BARLAS, Research Centre on Interactive Media, Smart Systems & Emerging Technologies, Cyprus

KYRIAKOS KYRIAKOU, Research Centre on Interactive Media, Smart Systems & Emerging Technologies, Cyprus

OLIVIA GUEST[*], Research Centre on Interactive Media, Smart Systems & Emerging Technologies, Cyprus

STYLIANI KLEANTHOUS[†], Open University of Cyprus, Cyprus

JAHNA OTTERBACHER[†], Open University of Cyprus, Cyprus

Machine-learned computer vision algorithms for tagging images are increasingly used by developers and researchers, having become popularized as easy-to-use "cognitive services." Yet these tools struggle with gender recognition, particularly when processing images of women, people of color and non-binary individuals. Socio-technical researchers have cited data bias as a key problem; training datasets often over-represent images of people and contexts that convey social stereotypes. The social psychology literature explains that people learn social stereotypes, in part, by observing others in particular roles and contexts, and can inadvertently learn to associate gender with scenes, occupations and activities. Thus, we study the extent to which image tagging algorithms mimic this phenomenon. We design a controlled experiment, to examine the interdependence between algorithmic recognition of context and the depicted person's gender. In the spirit of *auditing* to understand *machine behaviors*, we create a highly controlled dataset of people images, imposed on gender-stereotyped backgrounds. Our methodology is reproducible and our code publicly available. Evaluating five proprietary algorithms, we find that in three, gender inference is hindered when a background is introduced. Of the two that "see" both backgrounds and gender, it is the one whose output is most consistent with human stereotyping processes that is superior in recognizing gender. We discuss the accuracy – fairness trade-off, as well as the importance of auditing black boxes in better understanding this double-edged sword.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing design and evaluation methods**; • **Social and professional topics**;

Additional Key Words and Phrases: Bias detection; Gender; Image Tagging Algorithms; Social Stereotypes

---

[*]Also with Experimental Psychology, University College London, UK.

[†]Also with Research Centre on Interactive Media, Smart Systems & Emerging Technologies, Cyprus.

Authors' addresses: Pınar Barlas, p.barlas@rise.org.cy, Research Centre on Interactive Media, Smart Systems & Emerging Technologies, Nicosia, Cyprus; Kyriakos Kyriakou, k.kyriakou@rise.org.cy, Research Centre on Interactive Media, Smart Systems & Emerging Technologies, Nicosia, Cyprus; Olivia Guest, o.guest@rise.org.cy, Research Centre on Interactive Media, Smart Systems & Emerging Technologies, Nicosia, Cyprus; Styliani Kleanthous, s.kleanthous@rise.org.cy, Cyprus Center for Algorithmic Transparency, Open University of Cyprus, Nicosia, Cyprus; Jahna Otterbacher, j.otterbacher@rise.org.cy, Cyprus Center for Algorithmic Transparency, Open University of Cyprus, Nicosia, Cyprus.

---

1 INTRODUCTION

There is an increasing volume of research on bias in algorithmic systems, given the harm they can cause to both individuals and marginalized social groups. Attention to the topic arguably stems from the growing influence of opaque — often proprietary — algorithms in our information ecosystem. Algorithmic systems and processes are often positioned as "power brokers" that are not always held accountable for their decisions and actions [30, 38]. Furthermore, such systems are now ubiquitous in our everyday lives, being delegated "everyday tasks" and operating largely autonomously, without human intervention [106]. Automated content analysis on images is a prime example of an "everyday task" that underlies many modern applications and platforms. In the current work, we focus on *proprietary image tagging algorithms* (ITAs), which provide automated content analysis on an input image, returning a set of descriptive tags.

In user-facing applications, automated image tagging is used extensively for personalizing the user experience, with *Forbes* claiming that the technology is transforming the retail industry.[1] For instance, e-stores can now infer a shopper's "personal style" through automatic recognition of the visual characteristics of items of interest.[2] In-store shopping can also be enhanced with image tagging, which can help in understanding what a customer is interested in and how to enhance their experience.[3] ITAs are also increasingly used in the sensitive domain of dating applications. Analogous to inferring a shopper's "personal style," as mentioned above, Hinge[4] uses a proprietary service, Clarifai, to track user behavior, determining the look of person images that the user views within the app, thus acting as a virtual match-maker.[5] Other clients of Clarifai are using the service to provide feedback to dating app users, such that they can craft the "perfect" profile image.[6]

Beyond their use in development, image tagging services have also become popular with researchers of social platforms. In particular, proprietary services have enabled researchers to conduct large-scale studies of social media behavior, to better understand the content users are likely to share [53] or how the properties of images correlate to user engagement [5]. Given the prominence of people-related images in profiles and posts, others are using tools specifically designed for facial and/or emotion recognition, to understand the traits of users who share certain types of media [28, 70] or to detect users likely to suffer from mental health issues [45].

It is important to recognize not only the importance — but also the potential sensitivity — of these types of analyses. In addition to the inferences on individuals' personalities and emotions, recent work has used image analysis to understand public health and lifestyle issues such as smoking and obesity [42, 65]. Some have even exploited ITAs to study indications of gang violence on social media [11]. Given that researchers are using computer vision techniques — many of them being proprietary tools — in sensitive contexts where people-related media are analyzed, it is important to ensure that these tools treat people fairly. Fortunately, the community is taking steps to understand the *social biases* inherent in computer vision tools, and their underlying causes.

1.1 Image tagging algorithms and social biases

To date, much of the work on social biases in computer vision has focused specifically on facial recognition, given the obviously sensitive nature of this technology. In contrast to the ITAs that we study, facial recognition systems are specifically designed to analyze images of people, for instance, by matching a new face to a previously-seen one, or inferring the gender of the person in an image.

---

[1] forbes.com/sites/forbesagencycouncil/2018/07/16/use-ai-to-create-a-more-personalized-profitable-customer-experience

[2] vue.ai/solutions/omnichannel-personalization

[3] catchoom.com/blog/image-recognition-enables-scan-to-shop-retail-experiences/

[4] hinge.co

[5] blog.clarifai.com/4-ways-ai-is-improving-dating-apps

[6] blog.clarifai.com/clarifai-featured-hack-use-ai-to-tune-up-your-online-dating-profile

Previous research has shown that such inferences may appear accurate for men with light skin, but that the error rates rise dramatically when analyzing images depicting other social groups, and in particular, women and/or people with dark skin [15, 44, 60, 62]. Facial recognition researchers have attempted to mitigate such biases by training algorithms that simultaneously infer age, ethnicity and gender; however, obtaining solid performance in unconstrained settings remains a significant technical challenge [26, 69].

There is also evidence that facial recognition performs best on cisgender individuals. Scheuerman and colleagues [93] evaluated the performance of facial recognition algorithms on images of transgender people. They concluded that the algorithms were generally unable to infer non-binary genders correctly. Given such results, some socio-technical researchers have taken the position that gender inferences using the physical form of a person are fundamentally incompatible with the sociological understanding of gender and the lived reality for many individuals [46, 59].

In contrast to facial recognition tools, ITAs are not specifically designed to analyze person images. Rather, ITAs resemble object recognition algorithms, returning "tags" or labels describing the concepts depicted in the input image. Thus, one might initially expect them to be less prone to social bias. However, part of the functionality of an ITA is recognizing visual patterns in the images, which may result in high error rates for patterns that do not appear frequently in the training dataset. For example, commercial object recognition algorithms have been found to have higher error rates for images of items found in lower income versus higher income households [27]. Similarly, it has been found that systems meant for use in autonomous vehicles — therefore specifically trained to recognize pedestrians — had higher error rates for darker-skinned versus lighter-skinned pedestrians [105].

It is important to note that while ITAs are not specifically designed to perform gender recognition, they often present multiple gendered concepts; for instance, in addition to tags describing gender directly (e.g., "boy," "woman"), many tags describe concepts typically associated with gender (e.g., "pretty" versus "handsome"). In our own previous work, we conducted audits on six proprietary ITAs, on a controlled set of people images, the Chicago Face Database,[7] with no background contexts and in which people were clothed and positioned in the same manner [67].[8] We found that all ITAs used gender-related concepts in describing images, and that they were significantly more likely to use gender tags incorrectly when describing images of women versus men, as well as for people of color. In short, our results were largely in line with those mentioned above, concerning the social biases exhibited by facial recognition algorithms.

1.2   Goals of the current work

Proprietary ITAs are unlikely to be trained on standardized sets of people images. Rather, real training datasets tend to capture the prevalent biases in our social world. Researchers are increasingly discussing *training data* as a source of gender bias in computer vision, reporting that certain activity labels often appear more frequently for images of one gender, reflecting a salient gender stereotype (e.g., "shopping" or "cooking" on images of women; "golfing" or "snowboarding" on images depicting men) [50, 108]. These authors caution that biases may be amplified during training, resulting in even higher frequencies in the output as compared to the training datasets. As mentioned, in ITAs, we expect higher error rates in recognizing patterns that are less frequent, as compared to those that appear more often in the training data. Thus, we expect to find that images that are congruent with gender stereotypes are easier for the ITAs to describe, and that it will be more difficult for an ITA to recognize concepts in an image – including a person's gender as well as

---

[7]https://chicagofaces.org/default/
[8]This dataset is described in detail in the Methodology section.

the context – when the image does not conform with prevalent gender stereotypes (e.g., an image of a man in a kindergarten classroom, or a woman in a mechanics garage).

Therefore, the goal of the current work is to build upon the controlled auditing approach developed in [67]; however, rather than simply detecting social bias in the ITAs' descriptions of people and their inferred gender, we aim to better understand *how context intersects with the ITAs' interpretations of people*. Drawing from the social psychology literature describing how and why humans gender-stereotype others [23], we ask whether "observing" an individual in a more/less gender stereotype-congruent context changes the way the ITA describes the person, and whether its accuracy in inferring gender is affected.

We structure the article as follows:

- In Section 2, we motivate the need to develop auditing procedures for ITAs offered as cognitive services, and we review auditing approaches described by socio-technical researchers.
- Section 3 grounds our study in the social psychology of gender stereotyping, enabling us to benchmark the ITAs' behavior against that of humans. It also provides a brief review of how stereotypes are expressed in training data.
- We provide details on our methodology, the dataset created and the analyses performed in Section 4.
- Section 5 presents our findings, while in Section 6 we discuss the implications, considering the trade-off between achieving human-like accuracy in ITAs, and treating the depicted persons fairly.

The main contribution of our work is two-fold. First, we develop a reproducible methodology for auditing proprietary ITAs, which allows us to manipulate the background context in which the person image is introduced. Our datasets and code are publicly available. Secondly, we provide evidence that there is a trade-off between an ITA's gender inference accuracy and fairness, with the best performing ITA being the one that most closely mimics human gender stereotyping processes.

2 AUDITING COGNITIVE SERVICES

Developers, designers and researchers, who are not experts in artificial intelligence (AI) or machine learning (ML), are increasingly interested in incorporating these tools into their work. In this section, we discuss a key enabling mechanism – *cognitive services*, also referred to as *AI as a Service (AIaaS)*. We briefly review these trends, motivating the need to develop auditing procedures for these opaque services. In particular, we argue that cognitive services can and should be compared to human behavior, in order to better understand their social biases and potential harms.

Using ML has traditionally been challenging for non-experts. For example, researchers studying the use of ML by HCI students [85] and UX designers [31] reported in both groups an enthusiasm for incorporating ML into designs and prototype systems. Yet even participants who were experienced programmers expressed difficulty in their understanding and use of ML. In addition, novices frequently expressed an expectation for ML tools to "just work" as a black box without needing to first invest a large amount of upfront effort [85]. They also described difficulty in "gather[ing] data for their own tests" of the tools (i.e., auditing the tools).

Cai and Guo surveyed professional software developers to better understand their motivations for using ML as well as the difficulties faced [17]. Interestingly, developers often reported an aspirational motivation for using ML tools rather than a direct need for them in projects. However, many also expressed a desire for ML to be "demystified." This last finding underscores the fact that, to be transparent, algorithmic tools must not only be open (i.e., not protected by trade secrets) and "white box" (i.e., interpretable from a technical point of view); users/developers must also have the skills necessary to understand them [16].

Generally, the fast-moving development environment and the tendency for improving software reusability [83, 91], means that developers are often looking for easy-to-install frameworks or tools with minimal cognitive overhead, in an affordable and accessible way [29, 95]. One way that industry has responded to the interest by non-experts in using ML and AI, is through the provision of *cognitive services*, algorithmic processes inspired by human cognition, for solving various narrow-AI problems such as sentiment analysis of text, translation of text from one language to another, or visual recognition of objects and concepts depicted in images and video. Microsoft, one of the leading providers of cognitive services, through its Azure platform, has described them as "democratiz[ing] AI by packaging it into discrete components that are easy for developers to use in their own apps."[9] Indeed, many start-ups are taking advantage of cognitive services, and not only to enhance the capabilities of their systems and apps. According to *Forbes*, aligning themselves with the AI industry can also be commercially beneficial to new companies.[10]

Yet cognitive services arguably do not provide the "demystification" desired by many users, nor are they "well-tested" or regulated by third parties independent of the provider. In fact, in the vision-based services provided for image analysis and tagging (ITAs), the developer is not provided with the set of content tags that may be used to describe an image input for analysis. In other words, when developers or researchers use the pre-trained models to analyze images, they are black boxes; one can only observe the input image and the output tags the algorithm has inferred as describing the image content. Many algorithmic systems used extensively in everyday life are effectively black boxes, from search engines to social media news feeds. However, in the case of cognitive services, there is a real possibility to perpetuate social biases on the large scale, when developers build on top of these services, or when researchers use them to infer characteristics of human behavior on the Web and social platforms. For this reason, there is a need for reproducible auditing approaches, in order to gauge and monitor the behaviors and use cases of these tools.

2.1 Detecting biases in cognitive services

Figure 1 depicts a high-level, generic architecture of an algorithmic process, which is offered as a cognitive service to users. In the figure, we observe the typical ML pipeline, consisting of an input to the system, a model that has been trained based on a set of observations from the respective domain, and output that is received by the user. Typically, in the case of a user-focused system or process (e.g., a search engine, a social media news feed), the user interprets and applies the output depending on the context/problem at hand, and may again interact with the system, introducing new input(s). However, in the case of cognitive services, the user is a developer or researcher, who collects the output, using it in the creation of a new system, process or analysis.

Following Orphanou et al. [79], the points at which bias can manifest and be detected, appear in circles in Figure 1. As mentioned, training data (D) are often a source of bias. Likewise, algorithmic processing can also introduce bias (M), e.g., if developers use a statistically biased estimator in the model [25]. Fairness constraints (F) may be introduced within the system by developers, such that one interpretation of fairness is prioritized over others [63]. However, researchers have emphasized that in the context of socio-technical systems, there is never a single interpretation on "what is fair" [78, 102]. Nonetheless, audits on algorithmic systems typically emphasize the notion of *group fairness*, which holds that advantaged and protected groups should be treated no differently than others [86]. In contrast, *individual fairness* concerns the notion of consistency, holding that similar individuals should receive similar treatment by the system [9]. In summary, these are sources of bias originating from *within the system*, which of course, are typically not visible to the user. In

---

[9]blogs.windows.com/buildingapps/2017/02/13/cognitive-services-apis-vision
[10]forbes.com/sites/parmyolson/2019/03/04/nearly-half-of-all-ai-startups-are-cashing-in-on-hype/
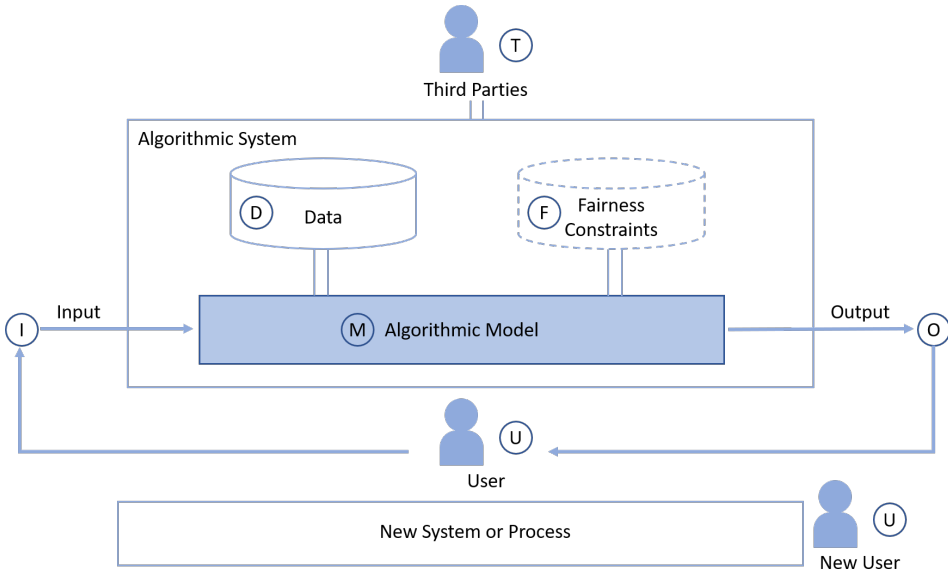
Fig. 1. System architecture: perpetuating bias with Cognitive Services

particular, in the case of a cognitive service, the "algorithmic system" is a black box, and the user observes only the input and the output.

Sources of bias originating *outside the system* include third party biases (T) as well as those introduced and/or detected by the user (U). Implicit and explicit constraints, given by third parties, may impact the design and performance of the algorithm and cause bias in the output. These include operators of the system, regulators, and other bodies which influence the use and outcomes of the system. A typical example is the modification of search engine results in order to comply with local laws in the country where the user is based.

Finally, when users interact directly with a system, they may contribute to bias in a number of ways, such as through the inappropriate use of the system or misinterpretation of system output (O). In the case of cognitive services, if the output reflects social biases, such as the ITAs' use of gender-related tags as previously mentioned, the biases will be perpetuated in the new system or process developed, potentially affecting new users. Therefore, while cognitive services provide access to state-of-the-art AI components, even for non-experts, there is a need to build in a layer of bias detection and mitigation. Next, we review the auditing approaches described in the literature, which can be used by stakeholders outside of the "black box," in order to scrutinize their behaviors.

2.2   Common auditing approaches

Auditing is increasingly used as a means to detect bias and possible discrimination in algorithmic systems. While formal auditing is typically done by developers with full access to the system, it is increasingly performed by other stakeholders, including third parties and even users. An example of a third-party audit can be found in the widely discussed 2016 ProPublica article, "Machine Bias" [3], concerning the proprietary COMPAS system, used in the U.S. justice system to predict the likelihood of recidivism. The authors, as data journalists assuming the role of *system auditors*, compared the data they collected from public criminal records concerning 10.000 defendants, to

the predictions made by the COMPAS system.[11] They cited systematic racial bias, such that the system's error rates were more than double for black versus white defendants.

Particularly in the case of proprietary systems, where full transparency (i.e., a code inspection) is impossible, Sandvig and colleagues [92] suggested auditing algorithms "from the outside." Such techniques are appropriate for auditing cognitive services, and we focus on summarizing relevant approaches that have been described in the literature. Eslami and colleagues [36] articulated two high-level approaches for auditing by third parties without access to the "insides" of the black box.

In a *within-platform audit*, the input is systematically manipulated, in order to study how the resulting output is affected. Many examples of this technique can be found in studies of search engine bias. Sweeney [100], in a test for racial bias in Google AdSense, conducted Web searches on names, manipulating them by their racial associations. She then analyzed the content of ads chosen by the algorithm and found that searches on names commonly given to Black children (e.g., DeShawn or Jermaine) were significantly more likely to be served up ads related to arrest, as compared to searches made on white-coded names (e.g, Emma, Jill). Kay et al. [57], Magno et al. [73], and Otterbacher et al. [82] all submitted queries to search engines to study the perpetuation of gender stereotypes in image search engines. Comparisons are made between properties of the input queries (e.g., the extent to which they describe an occupation or trait that is stereotypically masculine/feminine), and the properties of the output (e.g., the gender distribution in the images retrieved). Another example of using auditing to detect bias in a search engine can be found in the work of Kilman-Silver et al. [64] who examined the influence of geolocation on Web search personalization. They collected and analyzed Google results for 240 queries over 30 days from 59 different GPS coordinates, detecting the systematic differences in output.

A second approach, *cross-platform auditing*, can be useful for detecting cases where a system is generally biased, i.e., biased outputs can be observed for any inputs, and not just for certain subsets of inputs with particular properties. In their audit of the hotel rating system at Booking.com, Eslami et al. [36] compared the ratings of a random set of hotels at Booking.com, to those on two other platforms. Kulshrestha et al. [66] proposed an auditing technique that gauges information bias on Twitter, by benchmarking its behaviors against those of search engines. Their auditing technique considers both the input and output bias. Input bias allows the researchers to understand what a user would see if shown a set of random items relevant to her query. Similarly, [76] quantifies search engine bias, by collecting a fair results set, consisting of the results on a given query across multiple engines. The behaviors of a given engine are benchmarked against this combined set.

Hannák and colleagues [47] apply both within- and between-platform auditing approaches, in their study of social biases in online marketplaces. They study the rankings of job candidates, by their gender and race, comparing also the ranking behaviors at Fiverr and TaskRabbit. In the current study of ITAs offered as cognitive services, we also apply both approaches. As will be explained in Section 4, applying the within-platform approach, we gauge how each ITA's descriptions of people images changes when we manipulate the background context in which the person is pictured. In addition, we make cross-platform comparisons, to see which ITAs perform better on recognizing both the background context as well as the depicted person's gender.

2.3 Auditing cognitive services: What would people do?

We take inspiration from a recent article in *Nature*, describing an emerging, interdisciplinary science of Machine Behavior [89]. The authors describe this new field as drawing from both the disciplines that concern the design and implementation of AI and/or ML-based systems, and those that concern the behaviors of biological agents. This is because modern ML-based systems exhibit

---

[11]propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

"hybrid human-machine behavior," as they have been trained on large volumes of data generated by humans, representing aspects of the social world.

Just as cognitive services are inspired by human behavior and cognition, we aim to develop an approach for auditing proprietary, black box ITAs, benchmarking their behaviors against those we would expect from humans. This is not because we believe that ITAs operate in a manner similar to a human when interpreting an image of a person; cognitive services are named as such because they are inspired by biological cognition [58]. Rather, we benchmark their behaviors against what we would expect humans to do, as we are concerned with possible social biases that might prove harmful in the social world. Also, since the behaviors of the ITAs will be interpreted by the human user, it makes sense to use a human baseline. In short, just as many computer vision engineers evaluate algorithms with respect to their degree of similarity to human cognition [6], we wish to evaluate the ITAs' social behaviors in a parallel manner. To lay the groundwork for such a process, Section 3 explores the nature and origins of gender stereotyping in humans, as well as their expression in media and finally, the data used to train algorithmic systems.

3 GENDER STEREOTYPES

To systematically examine the ways in which ITAs describe images of people depicted in various social contexts, we first ask how we would expect a human to interpret the respective images. In particular, we review literature from social psychology, to understand the role of gender stereotypes in human behaviors. Finally, we briefly explore how gender stereotypes may be expressed in the data that is used to train algorithmic systems. Based on the review, we then articulate four research questions, which guide our analysis.

3.1 Possible origins and functions

Gender stereotypes have long been studied by social psychologists, who propose that people develop gender role beliefs through individual experiences [104]. For instance, if we repeatedly observe a certain gender in a particular role, this may lead us to infer that this gender is more suited for the role. At the societal level, these associations can become collective knowledge – stereotypes – influencing the way we talk to, think of, and view other people [101]. People often use stereotypes to justify a behavior or action and to define group boundaries [48]. Social stereotypes can also become activated in situations where a decision needs to be made and we lack important information; however, they can also prove harmful when inaccurate [12, 41].

Socialization facilitates performances, by encouraging individuals of a given gender to develop certain personality traits and skills [35, 96]. According to social role theory, there are three main concepts of social roles: *i*) Social position — where one stands in the social system (e.g., husband/wife); *ii*) Role expectation — how one is expected to behave according to their social position; *iii*) Role enactment — measuring how well one is performing according to role expectations [1]. Extending through these concepts, people make assumptions related to gender, the social role/s that each gender can assume, and how well one can perform a role.

3.2 Stereotypes in occupations and physical contexts

Gender stereotypes concerning occupations, which tend to be well-defined and prevalent in society, constitute a good example of the strong associations between gender and social role expectations. For example, it is widely accepted that socially desirable traits for men reflect competence, while traits for women reflect warmth [23, 40, 96]. As might be expected, the "warm women, competent men" stereotype has many implications in occupational contexts [24]. In particular, experimental results have shown that people generally expect men to be employed in high status jobs as compared

to women, and also expect men to be more influential in their respective positions [34]. Furthermore, gender stereotypes have been used to explain the division of labor in employment. Specific employment roles that have traditionally been dominated by men are thought to require an agentic personality and/or strong physical attributes, while employment roles that are women-dominated are typically believed to require more warm, feminine attributes [18, 104].

Likewise, gender can become associated with a physical context or an activity. For instance, research conducted within the school environment considered the gendered perceptions of both adults and young children. Specifically, the indoor physical environment, including rooms/zones, as well as children's play practices and activities, were often associated with a certain gender [14]. However, it has been suggested that activity-based stereotypes more often concern expectations for boys, whereas girls are more often the subject of appearance-based stereotypes [43].

In short, stereotypes affect the way people interpret others, relative to specific roles within society, as well as the way they expect others to behave. This can lead to being judgmental of those who follow a less stereotypical path. For example, we are familiar with the stereotype of the "stay-at-home mom." However, the growing concept of the "stay-at-home dad" is often unexpected, or even challenged, leaving men adopters of this role feeling embarrassed, if their role is revealed to the outside world, e.g., via social media [2]. Many studies describe how gender stereotypes can negatively impact children developmentally. For instance, in one study, girls appeared to be slower in a spatial task (stereotyped as a male strength) when their gender identity was activated compared to girls of the same age whose identity was not activated [94]. Such experiences can leave children feeling intimidated by their peers. Shenouda et al. emphasize how, even in contexts where boys and girls perform similarly in science and technology subjects, girls often grow out of those interests and occupations, having been negatively affected by gender stereotypes [94].

3.3 Gender stereotypes in media and technology

It is notable that gender stereotypes have been around for a long while [33], often resulting in discrimination [49] and inequality [18]. One might expect that in a modern world, where women are increasingly encouraged to pursue higher education and high-status careers, that we would have made more progress. There is a decades-long bibliography surrounding the perpetuation of gender stereotypes in the media, from television [21, 68, 75] and films [55], to newspapers [4, 52, 61], and even comics [19]. In a similar spirit, researchers are now questioning the role of the latest, AI-driven technology, in the perpetuation, and sometimes even the amplification, of social stereotypes. Certainly, these influences are hindering society's progress in gender equality.

In [97], the authors explored how stereotypes of strongly gendered professions (librarian, nurse, computer programmer, civil engineer) are expressed across digital platforms (Twitter, the New York Times online, Wikipedia, and Shutterstock). They found that gender stereotypes are most likely to be challenged by human creators and curators rather than algorithmic-dependent approaches, which showed little inclination towards breaking stereotypes. Specifically, the authors analyzed the context of the collected images, finding that women are largely underrepresented in images on digital platforms. Furthermore, this finding was consistent for both male- and female-dominated professions. In a similar vein, Kay et al. [57] found evidence of stereotype amplification in Google search results on profession-related queries. They also found that search results were rated higher by study participants, when they reflected stereotypes about careers. Considering the societal impact of these representations of the professions, they also demonstrated that shifting the representation of gender in image search results can influence people's perceptions of real-world gender distributions.

3.4   Stereotypes in training data

Developing computer vision algorithms such as ITAs typically relies on training data in which images have been labeled by humans. Datasets in which natural language is used to collect/encode image descriptions, have been found to be particularly susceptible to socio-cognitive biases [90]. In an analysis of Flickr30K, van Miltenburg found that the crowdworkers hired to annotate the data made various inferences on images depicting people, which do not logically follow from the image content. For instance, he noted cases of gender, racial and ethnic stereotypes, as well as other "unwanted inferences" (e.g., ethnicity marking, suggesting that images of white people are the default) [101]. Otterbacher et al. found similar patterns when crowdworkers were asked to describe highly controlled and uniform headshots of people of various races and genders [81]. Such findings on image datasets resonate with work suggesting that generally, natural language data (e.g., texts collected from the Web) are subject to reporting bias. Co-occurrences reported in text often do not respond to their frequency in the offline world, as people may over- or under-report something unusual or counter-stereotypical (e.g., a "male nurse") [13].

Blatantly racist or sexist language is typically less of a problem in curated training datasets, as it can be detected and removed using automated methods [84]. However, there are more subtle ways of expressing stereotypes, which are less likely to be noticed. In particular, in descriptions of images depicting people, stereotypes might be expressed through *linguistic biases*, a systematic asymmetry in the way that one uses language, as a function of the social group of the person(s) being described [8]. This bias is believed to be of a cognitive nature, since prototypes or stereotypical scenes are simply easier for humans to process [107].

When describing people (e.g., in an image), the use of more abstract versus concrete language conveys information about the observer's social expectations. The more expectancy-incongruent a person and her behavior appear to us, the more likely we are to describe the person with more concrete language, without extrapolating beyond the particular context [72]. In contrast, the more comfortable we are with the person, in the sense that she does not violate our expectations, the more likely we are to use more abstract language, making causal inferences about her identity, disposition or behaviors [37]. For example, given an image of a man wearing all white clothing, a concrete description might simply state "man in white," whereas "doctor" would be abstract.

Linguistic biases can be observed in tasks where participants describe images of people. Linguistic Expectancy Bias (LEB) predicts that we describe counter-stereotypical people more concretely as compared to more stereotypical people [8]. Stereotypical people, whose appearance is more expected, tend to be described abstractly, with inferences made by the annotator. This is also true of in-group members, who are more familiar to us. Evidence of LEB has been documented in the ESP Game dataset [80], in Flickr30K [101] as well as in Wikipedia-sourced data [103]. Continuing the previous example of an image of a man in white, LEB would predict that the depicted person's race might be correlated to properties of the descriptions of the image, with an image of a white man being more likely to receive the abstract, "doctor" description, as compared to a Black man, who would be more likely to be described more concretely.[12]

3.5   Research Questions

The previous findings demonstrate the importance of observing people in their contexts (e.g., social roles, occupations, activities) in the formation of gender stereotypes. Given this, we pose four research questions aimed at understanding the expression of gender stereotypes by ITAs:

- RQ1: Do ITAs "see" more women/men in a stereotypically feminine/masculine context?
- RQ2: Do ITAs recognize the background contexts, independently of the gender perceived?

---

[12]Our example assumes a North American cultural context.

- **RQ3:** Given that an ITA recognizes the background, are gender inferences more accurate in a stereotype-congruent background?
- **RQ4:** Considering the ITAs that show consistency on gender and context recognition, how do the image descriptions change when a background is introduced, and are these changes uniform across social groups?

To answer our questions, we develop a reproducible methodology, in which we impose a highly controlled set of people images onto various backgrounds. As will be explained, this allows us to make both between-image and within-image comparisons, to better understand how ITAs perceive the genders of people of various racial groups, in a number of contexts. Such comparisons were not possible in previous work [93], in which the images studied were collected from the wild (i.e., neither the person characteristics, nor the backgrounds could be controlled).

4 METHODOLOGY

4.1 Selecting contexts and images

We selected contexts associated with stereotypical jobs and social roles, based on labor statistics regarding gender,[13] and occupational stereotypes from previous literature. We then searched for images (on royalty-free image platforms such as Pexels[14] and Pixabay[15]) which could be used to represent these contexts, paying attention to their composition. Specifically, the images needed to be without people (to ensure the person-related tags would only be relevant to the person superimposed) and have a balanced layout (so that obscuring the middle of the image by imposing the picture of a person did not make it difficult to infer the context). In the end, we identified eight contexts — four that are stereotyped to be feminine and four that are masculine — with one image for each context as listed in Table 1. Examples of context images can be seen in Figures 2 and 3.

Table 1. List of contexts and the crowdworker perception of the gender stereotype for each.

Stereotypically Feminine	Median Score	Stereotypically Masculine	Median Score
F1: Wedding Dress Shop	0	**M1:** Mechanics Garage	9
F2: Baby Nursery	4	**M2:** Professional Kitchen	7
F3: Home Kitchen	4	**M3:** Sports Bar	7
F4: Kindergarten Classroom	5	**M4:** University Classroom	5

4.2 Understanding contexts and their perceptions

A crowdsourcing task was set up to understand the gender stereotypes associated with the images selected and their intensity. After a 24-response pilot (not included in our analyses), each of the eight images was presented to 50 crowdworkers for a total of 400 responses from 273 unique workers registered in the U.S. on the Amazon Mechanical Turk platform. The task asked the workers to consider the image, and then to (i) enter free text to answer the question "In which context was this image taken?" (ii) indicate on a sliding scale, "How masculine/feminine" they found the image (0: feminine, 5: neutral, 10: masculine), and (iii) declare their gender from a list of choices: male, female, non-binary/other, and prefer not to say. See Figure 2 for a screenshot of the task. The median time taken to complete the task was 121 seconds, while the payment was 1.00 USD per response.

---

[13]bls.gov/cps/cpsaat11
[14]pexels.com
[15]pixabay.com

Fig. 2. Screenshot of the crowdsourcing task for the image of a Baby Nursery.

The free text responses to the question regarding the location of the context was manually categorized into three groups – (i) the responses that described the context correctly, (ii) those that got the context completely wrong, and (iii) those that mentioned some relevant concepts but not all and therefore, got the context partially correct (for examples see Table 2). The responses that got the context completely wrong were excluded, leaving 374 responses that we analyzed. Using these remaining responses, we calculated the median score given by the crowdworkers. One image got a median score of 0 (highly feminine) and another a median of 9 (highly masculine). The median scores of the rest of the images can be seen in Table 1. It was a (pleasant) surprise that two contexts (F4, M4) were viewed as gender-neutral by workers, despite their traditional stereotypes.

Table 2. Responses to the first question in the crowdsourcing task.

Image context	Response	Relevance coding
Home Kitchen	"kitchen in house"	Correct
Baby Nursery	"bedroom"	Partially correct
Mechanics Garage	"zoo"	Incorrect

Table 3. Number of CFD (baseline) images by the person's race and gender.

	Asian	Black	Latinx	White	Total
Women	57	104	56	90	307
Men	52	93	52	93	290
Total	109	197	108	183	597

Fig. 3. Background F1 (left), LM-201 from the CFD (center), and the composite image (right).

Table 4. Output tags from Watson for the example images in Figure 3.

Background image F1	CFD Image LM-201	LM-201 in Context (F1)
dressing_room, indoors, clothing_store, shop, retail_store, building, cloakroom, dress_rack, rack, support, closet, clothing, fabric, reddish_orange_color, grey_color	soul_patch_facial, hair, stubble, actor, person, young_man, juvenile_person, ash_grey_color, light_brown_color	person, dressing_room, indoors, mantelet_(women's_cape), cloak, overgarment, garment, mantilla_(women's_scarf), scarf, clothing, fabric, ivory_color, ash_grey_color

4.3 Original person images and creating composite images

Created primarily for use as stimuli in psychology studies, the Chicago Face Database (CFD) is a dataset of highly-controlled images of people from various races and genders [71]. Each individual is wearing the same grey t-shirt, standing in front of the same white background, looking straight at the camera, under the same lighting conditions. The images were post-processed to ensure the faces were the same size. Each person was asked to pose with a neutral expression; while the CFD has additional images of some of the individuals (portraying certain emotions), we focused only on the subset of 597 images from v2.0.3 of the CFD[16] where the person depicted had a neutral expression. Self-reported demographics (from mutually exclusive groups of four races and of two genders) are detailed in Table 3, while an example image from the CFD can be seen in Figure 3.

We used a script on Adobe PhotoShop to remove the white background pixels from the CFD person images, by using the "Magic Eraser Tool" on the top left corner and the top right corner of each image. We manually checked resulting images for quality control and used the tool again to remove any remaining white backgrounds.

Next, we used an additional Python script, which uses the Pillow Library,[17] to first normalize the background images (Table 1) by scaling and then cropping so that the images had identical width and height. We then had the code read in the 597 CFD images with the transparent backgrounds and superimpose them onto the new, normalized background images. The script aligned the bottom of the two images, and horizontally centered the CFD person image. One example of the created composite images can be seen in Figure 3. The code, which also allows for automatic scaling to

---

[16]chicagofaces.org
[17]pypi.org/project/Pillow

ensure a standard ratio of CFD image pixels to unobscured background image pixels, has been released publicly along with our dataset.[18,19] In total, we had 5,381 images: the 597 original CFD images to use as a baseline, 4,776 composite images from superimposing the CFD images onto eight different backgrounds (contexts), and the eight background images themselves.

4.4 Collecting and clustering descriptive tags

We collected responses from ITAs on the images across our dataset. In particular, we used five general, pre-trained models for tagging offered as cognitive services by the following providers:

- Amazon Rekognition Image[20] (hereon: Amazon, A)
- Clarifai[21] (hereon: Clarifai, C)
- Imagga Auto-tagging[22] (hereon: Imagga, I)
- IBM Watson Visual Recognition[23] (hereon: Watson, W)
- Microsoft Computer Vision[24] (hereon: Microsoft, M)

We executed a series of RESTful calls to upload each image (within each of the eight background contexts as well as the CFD originals with no background) into each of the five services using HTTP Requests and saved their response as the output of this process. Because the five ITAs use different formatting for their output and do not follow the same structural guidelines, we did some pre-processing on the data to get the raw tags in a similarly-formatted output. In particular, we tokenized the tags and where a tag had more than one word, we substituted the spaces with underscores to handle the phrase as one tag. Examples of tags can be seen in Table 4.

In order to efficiently compare the different tags used by the ITAs – where a single concept may be represented by different tags across taggers – e.g. "eye," "eyes" — we created a typology that maps the tags to a set of common themes. We applied an inductive thematic analysis [51] to the tag sets, creating "clusters" which housed concepts linked together. For further higher-level analysis, we grouped these smaller "sub-clusters" into four large "super-clusters". The clusters relevant to our current analysis are described in Table 5. It should be noted that the names of the clusters are simplified for convenience into single-word titles, but actually refer to broader, related concepts as well, and that the clusters are not mutually exclusive; tags may fall under more than one cluster (e.g., "girl" falls under both the "Feminine" and "Age" sub-clusters). It should also be noted that not all ITAs use all clusters of tags. In particular, Amazon does not use any *abstract* tags.

A detailed motivation for this process, as well as information on the clusters not mentioned here, can be found in [7]. It should be noted that our dataset, along with the typology of tags and the dictionaries of the thematic clusters, is a resource that our team is continually updating. The typology was first developed in 2018, based on machine and human descriptions of the baseline CFD images. The analyses reported in the current work were performed on the dataset of composite (i.e., CFD images on the background contexts) images, which was collected in December 2019, when our typology of tags was also updated.

4.5 Analysis

4.5.1 Mapping output to thematic clusters. We used the descriptive tags received from five ITAs for the 5,381 images and the relevant tag dictionaries from our typology for the analysis, as detailed in

---

[18]github.com/oliviaguest/CFD-backgrounds

[19]doi.org/10.7910/DVN/4W5GOW

[20]aws.amazon.com/rekognition/image-features/

[21]clarifai.com/developer/guide/

[22]imagga.com/solutions/auto-tagging

[23]ibm.com/watson/services/visual-recognition/

[24]azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

Table 5. Selection of relevant thematic clusters: names, explanations, & example tags.

Cluster	Description	Examples
Demographics	*Tags that describe the inferred gender, age and/or origin(s) of the depicted person*	
Masculine	Tags that refer to a masculine gender identity or expression	man, masculinity
Feminine	Tags that refer to a feminine gender identity or expression	girl, hostess
Concrete	*Tags that describe directly observable attributes of the image or the depicted person*	
Abstract	*Tags that describe the inferred, subjective, or conceptual attributes of the person*	
Judgement	Tags that describe an opinion or subjective description	fine-looking, strange
Traits	Tags that refer to a personality trait or enduring characteristic	funny, serious
Emotion	Tags that refer to an emotional, mental, or temporary physical state	joy, pain
Occupation	Tags that refer to a job, a field of work, or a social role	chef, gang
Other	*Tags that do not directly fit into any of the previous clusters*	

Table 6. List of contexts, example tags & dictionary size.

Context	Example tags	Dictionary size
F1: Wedding Dress Shop	bride, groom, shopping, veil, wedding_dress	30
F2: Baby Nursery	crib, family, home, home_decor, nursery	7
F3: Home Kitchen	cooktop, domestic, family, kitchenware	24
F4: Kindergarten Classroom	elementary_school, kindergarten, toy	12
M1: Mechanics Garage	auto_mechanic, car_repair, garage, machine	92
M2: Professional Kitchen	chef, major_appliance, restaurant_kitchen	18
M3: Sports Bar	bar, barmaid, beer, pub, public_house, tavern	25
M4: University Classroom	auditorium, graduation, scholar, university	23

Table 5. In addition, we created a dictionary of related tags, observed across all five services, which describe each of the eight background contexts, as shown in Table 6.

The output for each image/ITA pair in our dataset was vectorized, in two ways; we calculated the proportion of tags in the description, per sub- and super-cluster. In other words, the super-cluster representation tells us how a given ITA "saw" the image, in terms of its use of demographic, concrete and abstract tags to describe it. In contrast, the sub-cluster vector tells us the extent to which the ITA used masculine and feminine tags in its description of the image, along with each of the four abstract tags. In addition, for the image/ITA pair, we also use the context dictionaries to calculate the proportion of output tags that refer to the respective context (e.g., for a wedding image, what proportion of tags describe "wedding" concepts).

4.5.2 Recognizing backgrounds and gender. With respect to gauging the tagging algorithms' recognition of the background context, we developed a high-recall method; if, for a given image, the tagger's output description contained at least one tag related to the background, we considered it to have "seen" the background. With respect to gender, when the description used more feminine than masculine tags, we consider the tagger to have inferred a woman, and vice versa. Otherwise, we consider the inference to be neutral.

4.5.3 Addressing the RQs. To answer each research question, we conducted appropriate quantitative and/or statistical analyses to understand the behaviors of the ITAs across background contexts. To answer RQ1, concerning what the algorithms "see" when CFD images are imposed into contexts, we simply count, for each ITA and context combination, the number of images inferred as depicting a person of each gender, as well as the images not gendered by the tagger. We then compare these counts to those observed when the tagger processes the baseline images.

RQ2 concerns whether or not there is evidence that a tagger's recognition of the background is dependent of its ability to recognize the depicted person's gender (and vice versa). Following Zhu and Gigerenzer [109], we use Bayes' theorem, comparing the marginal probability of a given tagger recognizing a background context, with the conditional probability of inferring the background, given that it has described the depicted person as being a woman, a man or neutral.

To make within-tagger/between-context and between-tagger comparisons on gender recognition accuracy (RQ3), we consider only the cases in which there is evidence the algorithm inferred (at least partially) the background context, using at least one related tag. We then calculate the F_1-measure [87] for recognizing images of men and women, respectively, by each of the five taggers, within each of the eight background contexts. We compare the tagger's performance to its respective F_1-measure on the baseline images.

Finally, RQ4 asks how the ITAs' descriptions of the images change, from the baseline to the cases where the stereotyped backgrounds are added. To this end, our analysis focuses on within-image (i.e., paired) comparisons, in which we compare the vectorized representation of a given baseline image, to its vectorized representation in a given background context. In particular, we use the cosine distance metric for our comparisons [98] evaluating *individual fairness*. The analysis focuses on the Clarifai and Amazon taggers since, as will be shown, these taggers recognize both the background contexts and gender more consistently as compared to the others. Within each ITA, we examine the extent to which image descriptions change when put into a background, and whether or not the changes are uniform across race and gender groups (that is, respecting *group fairness*).

5 FINDINGS

5.1 Do ITAs "see" more women/men in a stereotypically feminine/masculine context?

RQ1 asks if there is a change in the gender inferences that taggers make when the background context changes, regardless of whether they are correct. Table 7 details the inferences made on the baseline (i.e., no background) images by each tagger, as well as the genders inferred in the composite images (i.e., counts of Women (W), Men (W) and Neutral (N) inferences). The cells in which the number of inferred women or men has increased, when the gender-congruent background is imposed, are *highlighted in grey*.

In most cases, taggers actually use fewer gender-related tags when describing the composite images, as compared to the baseline. In other words, the descriptions of the images in context become more *gender neutral*. The one clear exception is the Clarifai tagger. Relative to baseline, the images with the four stereotypically feminine backgrounds are more often described with feminine tags (i.e., inferred as depicting women). Similarly, on three of the masculine backgrounds (M1, M2, M4), Amazon "sees" more men as compared to baseline. For the remaining taggers (I, M, W), the

Table 7. Number of images inferred as depicting a [Woman - Man - Neutral]. Ground truth is [307 - 290 - 0]. The highlighted cells indicate that the number of inferred women or men has increased from the baseline.

	Amazon	Clarifai	Imagga	Microsoft	Watson
Baseline.	329 - 67 - 201	128 - 359 - 110	1 - 465 - 131	13 - 434 - 150	127 - 141 - 329
F1.	9 - 46 - 542	370 - 115 - 112	1 - 238 - 358	75 - 109 - 413	0 - 0 - 597
F2.	40 - 148 - 409	217 - 325 - 55	0 - 434 - 163	35 - 354 - 208	60 - 86 - 451
F3.	18 - 73 - 506	191 - 334 - 72	0 - 182 - 415	99 - 194 - 304	0 - 0 - 597
F4.	0 - 21 - 576	164 - 412 - 21	0 - 160 - 437	50 - 459 - 88	4 - 0 - 593
M1.	0 - 208 - 389	17 - 404 - 176	0 - 446 - 151	2 - 451 - 144	0 - 0 - 597
M2.	5 - 177 - 415	57 - 342 - 198	0 - 359 - 238	84 - 328 - 185	0 - 0 - 597
M3.	31 - 43 - 523	250 - 307 - 40	0 - 184 - 413	71 - 402 - 124	0 - 0 - 597
M4.	126 - 348 - 123	173 - 283 - 141	0 - 387 - 210	19 - 416 - 162	6 - 5 - 586

composite images are more likely to be described in a gender-neutral manner, although Microsoft does "see" more women, as compared to baseline, in all of the feminine contexts.

5.2 Do ITAs recognize the background contexts, independently of the gender perceived?

We first calculate the marginal probability, for each tagger/background combination, of the tagger having recognized the background. We also calculate the conditional probability of "seeing" the background, given that the tagger has described the depicted person as being a woman, a man, or neutral. It should be noted that the tagger could be inaccurate concerning the depicted person's gender; here, we only consider the manner in which the person is interpreted by the tagger. Note that cells marked with "NA" indicate cases where no women/men were "seen" by the tagger.

If the recognition of the background and the person's gender are independent events, the marginal and conditional probabilities should be equal to one another. This is clearly the case with Clarifai, where the backgrounds are always recognized, regardless of whether a man or woman is perceived.

Table 8. Prop. of images where background ("Bg") is recognized (marginal vs. conditional probabilities).

	Amazon	Clarifai	Imagga	Microsoft	Watson
F1. $Pr(Bg)$	1.000	1.000	0.005	0.434	0.987
$Pr(Bg|W)$	1.000	1.000	0	0.347	NA
$Pr(Bg|M)$	1.000	1.000	0	0.450	NA
$Pr(Bg|N)$	1.000	1.000	0.008	0.450	0.987
F2. $Pr(Bg)$	0.997	0.983	0	0	0.008
$Pr(Bg|W)$	1.000	1.000	NA	0	0
$Pr(Bg|M)$	1.000	0.969	0	0	0
$Pr(Bg|N)$	0.995	1.000	0	0	0.011
F3. $Pr(Bg)$	1.000	1.000	1.000	1.000	1.000
$Pr(Bg|W)$	1.000	1.000	NA	1.000	NA
$Pr(Bg|M)$	1.000	1.000	1.000	1.000	NA
$Pr(Bg|N)$	1.000	1.000	1.000	1.000	1.000
F4. $Pr(Bg)$	0.812	1.000	0	0.672	0
$Pr(Bg|W)$	NA	1.000	NA	0.820	0
$Pr(Bg|M)$	0.857	1.000	0	0.673	NA
$Pr(Bg|N)$	0.811	1.000	0	0.580	0
M1. $Pr(Bg)$	1.000	1.000	0.318	1.000	1.000
$Pr(Bg|W)$	NA	1.000	NA	1.000	NA
$Pr(Bg|M)$	1.000	1.000	0.202	1.000	NA
$Pr(Bg|N)$	1.000	1.000	0.662	1.000	1.000
M2. $Pr(Bg)$	0.194	1.000	0.062	0.203	0.862
$Pr(Bg|W)$	0.400	1.000	NA	0.238	NA
$Pr(Bg|M)$	0.209	1.000	0.003	0.177	NA
$Pr(Bg|N)$	0.186	1.000	0.151	0.232	0.862
M3. $Pr(Bg)$	0.893	1.000	0.888	0	1.000
$Pr(Bg|W)$	0.968	1.000	NA	0	NA
$Pr(Bg|M)$	1.000	1.000	1.000	0	NA
$Pr(Bg|N)$	0.879	1.000	0.838	0	1.000
M4. $Pr(Bg)$	0.921	1.000	0.879	0	0.817
$Pr(Bg|W)$	0.937	1.000	NA	0	0.667
$Pr(Bg|M)$	0.931	1.000	0.845	0	0.800
$Pr(Bg|N)$	0.878	1.000	0.943	0	0.819

Although this is not always the case for Amazon, there is one interesting observation. When Amazon "sees" a gendered person, the likelihood of recognizing the background is higher, as compared to the cases where its description is gender neutral (see F2, F4, M2, M3, M4, in Table 7).

The other three taggers (I, M, W, in Table 8) are more idiosyncratic in their behaviors. For instance, when Imagga recognizes the background relatively accurately (e.g., F3, M3, M4), it continues (as in the baseline) to infer nearly all depicted people as men. When Microsoft fails to recognize the background (e.g., F2, M3, M4) its behavior is like that exhibited in the baseline — it infers most images as depicting men. Finally, overall, when Watson recognizes the background accurately (e.g., F1, F3, M2, M3), it makes few inferences about gender. Interestingly, in the background it did not recognize (F2), it made the most gender inferences (see Table 7).

5.3 Are gender inferences more accurate in a stereotype-congruent background?

We used F_1-measure to consider the accuracy of each tagger in using gender-appropriate tags on images of men (left side of Table 9) and women (right side) on the baseline images, as well as the eight composite images. Here, we restricted the analysis to the set of images upon which the background was recognized. Cells with "NA" indicate cases where the tagger did not recognize the background in at least 10% of the images. Shaded cells indicate cases where we observe an increase on F_1-measure, where a background context that is stereotype-congruent has been imposed.

Considering the images of men, we observe that the effects of adding a stereotypically masculine background on the gender inference accuracy, are mixed. Imagga, Microsoft and Watson clearly struggle to recognize both backgrounds and gender. Only Clarifai (M2, M3, M4) and Amazon (M1, M2, M4) demonstrated improved F_1 measure over the baseline. However, it can be noted that adding the incongruent (i.e., feminine) background, actually improves the recognition of men in depicted images, for two cases for Amazon and Clarifai (F2, F3).

Examining the right side of Table 9 reveals that adding a stereotypically feminine background only helps Clarifai recognize women. We observe that Imagga, Microsoft and Watson struggle to recognize both backgrounds and gender. Amazon recognizes the backgrounds but F_1 falls across all eight composite images. Finally, overall, gender inference accuracy is higher on images of men versus women within each background/tagger combination. The only exception is that of Clarifai in context F1, where F_1 for women is higher (0.88 versus 0.57).

Table 9. F_1-measure on gender tagging: images of men (left) and women (right). Shaded cells indicate an increase on F_1-measure as compared to the baseline.

	Images of Men					Images of Women				
	A	**C**	**I**	**M**	**W**	**A**	**C**	**I**	**M**	**W**
Baseline	.37	.89	.75	.79	.56	.84	.59	.01	.08	.57
F1: Wedding	.27	.57	0	.41	0	.06	.88	0	.42	0
F2: Nursery	.68	.92	NA	NA	0	.23	.82	NA	NA	0
F3: Kitchen	.39	.91	.69	.68	0	.11	.76	0	.49	0
F4: Kindergarten	.14	.82	NA	.71	NA	0	.70	NA	.32	0
M1: Mechanic	.82	.83	.48	.72	0	0	.10	0	.01	0
M2: Prof kitchen	.75	.91	0	.63	0	.07	.31	NA	.39	0
M3: Sports bar	.28	.91	.68	NA	0	.20	.88	0	NA	0
M4: University	.88	.95	.76	NA	.02	.59	.72	0	NA	.03

5.4 How do descriptions of images-in-context compare to those of the baseline, CFD images?

Next, we consider more in depth the behaviors of the two ITAs — Amazon and Clarifai — that were observed to have the most consistency on recognizing the backgrounds, as well as on using gender-related tags appropriately, as compared to the other ITAs. Tables 10 and 12 detail the mean cosine distance between the original CFD image and the composite image for a given background, within each of the eight social groups (i.e., gender/race combination, *where A: Asian, B: Black, L: Latinx, W: White and W: Women, M: Men*). For each background context, an ANOVA was conducted, with main effects on the gender and race of the depicted person, as well as an interaction term. The right-most column in each table shows the result of Tukey's Honestly Significant Differences (HSD) test. Applying the Bonferroni correction for post-hoc comparisons, we consider differences with a p-value less than 0.006 to be significant.

It is important to keep in mind that all ITAs output a limited number of tags; therefore, it is expected that when a CFD image is imposed onto a background, fewer tags are "available" to describe the person, since some should be describing the context. We use the vectorized representation of each image along with the cosine distance metric to compare how ITAs describe images in different contexts. Specifically, we make *within-image* comparisons (i.e,. same person image, context vs. baseline) and *between-image* comparisons within the same context (e.g., men vs. women wedding images). We avoid making comparisons between contexts as we cannot be sure of the number of available tags that are relevant to each context (i.e., some contexts might have more available tags than others). In other words, we are simply answering the question, *"Do some social groups' interpretations by ITAs change more than others, when the image is put into context?"*.

5.4.1 Clarifai. In the Nursery context (F2), Clarifai's interpretations of the images do not differ greatly from its baseline interpretations (i.e., cosine differences are all less than 0.10). In addition, there are no significant differences between social groups in terms of the mean cosine distances. However, for the other seven backgrounds, the extent to which Clarifai's interpretations change is correlated to social group. It is of note that for the backgrounds on which there is a significant gender effect (F1, F3, M1), it is always the case that the interpretations of images depicting women change more, when introducing a background context, than those of men. Likewise, where there is a significant effect on race (F1, F4, M1, M2, M3, M4), is it always the case that Clarifai's interpretations of images of Black people change more than those depicting other racial groups.

Given that the *wedding* (F1) and *mechanic* (M1) settings were interpreted as being *highly gendered* in our study with crowdworkers, we consider these two cases in greater detail. While Table 10

Table 10. Clarifai - Mean cosine distance between CFD and background images and results of ANOVA with Tukey HSD test.

	AW	BW	LW	WW	AM	BM	LM	WM	Sig. Differences
F1: Wedding	0.074	0.106	0.052	0.043	0.034	0.082	0.026	0.032	Gender: F > M
									Race: B > A, L, W
F2: Nursery	0.086	0.091	0.078	0.080	0.077	0.092	0.068	0.071	n.s.
F3: Kitchen	0.211	0.200	0.186	0.192	0.146	0.142	0.143	0.166	Gender: F > M
F4: Kindergarten	0.074	0.106	0.056	0.055	0.071	0.103	0.059	0.073	Race: B > A, L, W
M1: Mechanic	0.201	0.291	0.244	0.224	0.197	0.264	0.224	0.204	Gender: F > M
									Race: B > A, L, W
M2: Prof kitchen	0.160	0.187	0.133	0.137	0.147	0.157	0.145	0.142	Race: B > W
M3: Sports bar	0.075	0.106	0.070	0.068	0.105	0.108	0.085	0.083	Race: B > L, W
M4: University	0.164	0.213	0.160	0.161	0.140	0.182	0.156	0.156	Race: B > A, L, W

Table 11. Clariai - Within-image and between-image analyses. Significant differences consistent with gender stereotypes are highlighted. ∗ p< .002

Context	Attribute	Within-image			Between-image		
		Women	Men	t	Women	Men	t
Wedding (F1)	**Demographics**	-0.1002	-0.0009	-9.72*	0.099	0.118	-7.43*
	Abstract	0.0217	0.0064	1.23	0.216	0.180	7.81*
	Concrete	0.136	0.166	-2.62*	0.345	0.372	-5.16*
Mechanic (M1)	**Demographics**	-0.0507	-0.0351	-4.29*	0.123	0.095	19.1*
	Abstract	-0.1423	-0.1302	-2.43	0.052	0.055	n.s.
	Concrete	0.0497	0.0300	3.57*	0.316	0.311	n.s.

presented the overall cosine distance between CFD and composite images (i.e., over all attributes), we now consider changes in the individual attributes: demographic, abstract and concrete tags from our typology. For each image, we compute the difference in the scores on a given attribute, i.e., $S_{composite}$ - $S_{baseline}$.

Table 11 shows, for each gender, the change in the use of each type of attribute when the image is placed into a context. Two analyses are detailed: within-image comparisons (context vs. CFD — left side of Table 11) and between-image comparisons (women vs. men in context — right side). Table 11 details the mean differences for the images of each gender, along with the Welch t-test to compare across genders. For the between-image comparisons, we simply compare the gender-group mean proportion on each attribute on the composite images, using the t-test to compare them.

What we observe is that in the "wedding" (F1) context, while Clarifai uses more concrete tags for both men and women, relative to baseline, the increase is less for women than men. Similarly, the use of demographic tags is less in the wedding context (relative to baseline) for both genders, with the decrease being greater for women. Finally, there is an increase in the use of abstract tags to describe wedding images of men and women, although the difference in increase between genders is not statistically significant. Significant differences that are consistent with gender stereotypes are highlighted in grey. Figure 4 further examines the within-image differences across race/gender.
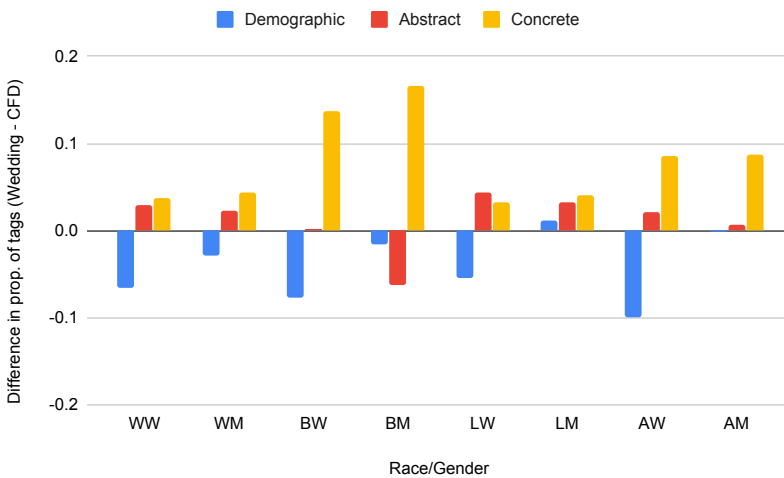

Fig. 4. Clarifai descriptions of wedding context: mean differences in use of tags relative to baseline.
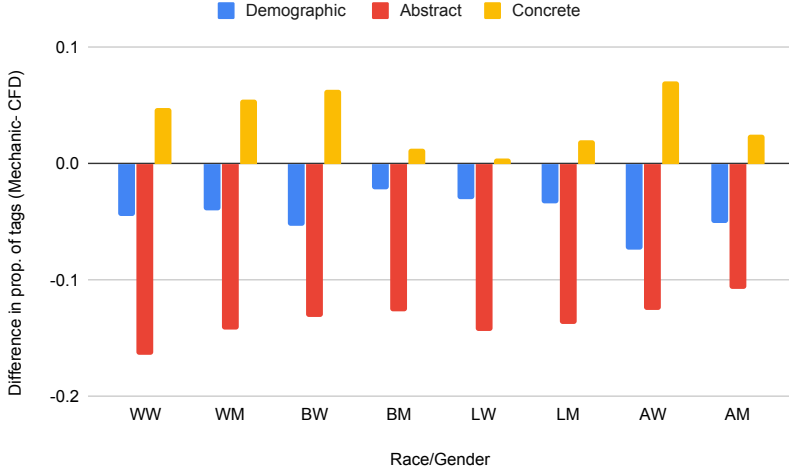
Fig. 5. Clarifai descriptions of mechanic context: mean differences in use of tags relative to baseline.

Here, we observe the effect of race noted previously. In particular, images of Black men in the wedding context become less abstract as compared to baseline. Thus, we see evidence that in general, *Clarifai finds women more congruent to the wedding context, and Black men less congruent to this context*. Further support for this is found in the between-image comparisons; here, we see that overall, descriptions of women in the wedding context use more abstract tags, while those of men use more demographic and concrete tags.

In the "mechanic" (M1) context, the changes relative to baseline are common to women and men: fewer demographic and abstract tags are used in Clarifai's descriptions, but more concrete tags are used. However, the magnitude of these changes differs, with respect to the use of demographic and concrete tags. Figure 5 breaks the differences out by race/gender. As found previously, *descriptions of Black men differ from the other groups*; while their descriptions do become less abstract/more concrete relative to baseline, the magnitude of change is less than the other groups. Also, the gender effect can be seen when race is held constant, e.g., descriptions of women become more concrete in the mechanic context, as compared to men, with the exception of BW/BM.

5.4.2 Amazon. Table 12 analyzes the global mean cosine distance between the composite and CFD images, per social group. Across all eight background contexts, there is a significant gender effect; *Amazon's descriptions of images of women change more than those of men*, relative to baseline. In addition, in all contexts other than kindergarten classroom (F4), there is a significant race effect; the *descriptions of images of Asian people change more than those of other races*, relative to baseline. In one context (Nursery, F2), this is also true of images of Black people, as compared to whites.

Table 13 takes a closer look at the wedding and mechanic contexts. In both contexts, we observe the same trend for women and men: images in context are described with fewer demographic tags and more concrete tags, although the magnitude of the change differs. In the between-image analysis, only one finding is consistent with gender stereotyping: more demographic tags are used to describe men in the wedding context versus women, who are stereotype congruent. Considering Figure 6, we observe that within each race, women receive more concrete descriptive tags than men, in comparison to their baseline images. However, men receive more demographic tags. Likewise, in Figure 7, WW, LW and AW receive more concrete tags as compared to men. However, across all

Table 12. Amazon - Mean cosine distance between CFD and background images and results of ANOVA with Tukey HSD test.

	AW	BW	LW	WW	AM	BM	LM	WM	Sig. Differences
F1: Wedding	0.099	0.078	0.085	0.087	0.073	0.067	0.0809	0.081	Gender: F > M Race: A > B
F2: Nursery	0.094	0.072	0.078	0.062	0.049	0.053	0.043	0.044	Gender: F > M Race: A, B > W
F3: Kitchen	0.149	0.119	0.134	0.111	0.109	0.113	0.100	0.108	Gender: F > M Race: A > B, W
F4: Kindergarten	0.081	0.067	0.069	0.062	0.031	0.037	0.025	0.032	Gender: F > M
M1: Mechanic	0.132	0.097	0.117	0.109	0.084	0.077	0.079	0.081	Gender: F > M Race: A > B, W
M2: Prof kitchen	0.105	0.073	0.086	0.085	0.069	0.063	0.064	0.077	Gender: F > M Race: A, W, > B
M3: Sports bar	0.136	0.101	0.105	0.094	0.080	0.080	0.062	0.070	Gender: F > M Race: A > B, L, W
M4: University	0.074	0.058	0.060	0.051	0.038	0.036	0.034	0.037	Gender: F > M Race: A > W

Table 13. Amazon - Within-image and between-image analyses. Significant differences consistent with gender stereotypes are highlighted. ∗ p< .002

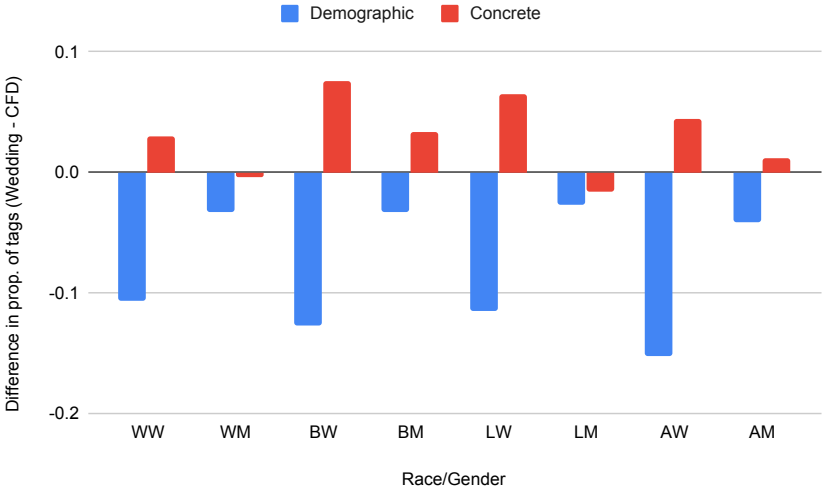| Context	Attribute	Within-image			Between-image		
		Women	Men	t	Women	Men	t
Wedding (F1)	**Demographics**	-0.123	-0.033	-18.5*	0.001	0.006	-5.21*
	Concrete	0.054	0.008	5.01*	0.537	0.496	5.40*
Mechanic (M1)	**Demographics**	-0.124	-0.016	-22.1*	0.0005	0.022	-24.3*
	Concrete	0.090	0.073	n.s.	0.574	0.561	n.s.

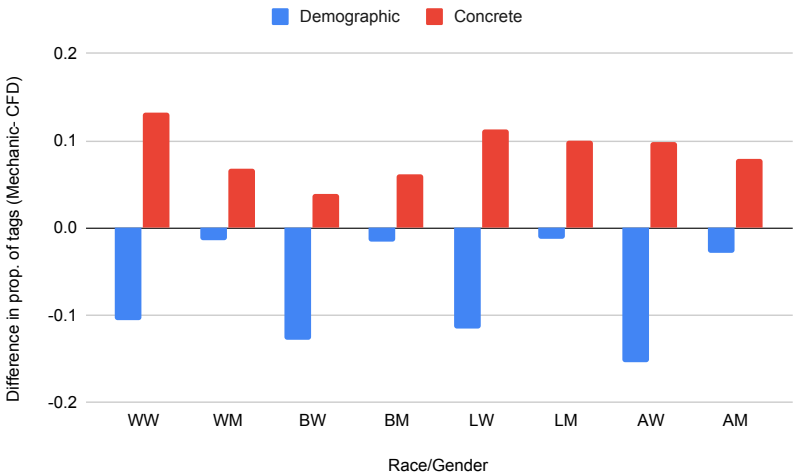Fig. 6. Amazon descriptions of wedding context: mean differences in use of tags relative to baseline.

Fig. 7. Amazon descriptions of mechanic context: mean differences in use of tags relative to baseline.

four races women receive fewer demographic tags than men, as compared to baseline, which is not consistent with social psychology theories of social stereotyping.

5.5 Summary of findings

Table 14 presents a concise summary of findings across the four research questions, with a focus on the two ITAs (Clarifai and Amazon), which were able to consistently recognize both the background and the gender of the person depicted. In the next section, we put the findings into the broader context, considering the accuracy versus fairness trade-off in ITAs.

Table 14. Summary of findings.

Research Question	Clarifai	Amazon
1. Do ITAs "see" more women/men in stereotypically feminine/masculine contexts?	Yes, for both W&M	Yes, for M only
2. Do ITAs "see" contexts, independently of the gender perceived?	Yes	No, context recognition is often better when gender is perceived.
3. Are gender inferences more accurate in a stereotype-congruent background?	Yes, for both W&M	Yes, for M only
4. How do descriptions of images-in-context compare to those of baseline images?	– Descriptions of W change more than those of M in strongly gendered contexts – Descriptions of Black people change more than others (6/8 contexts) – Use of attributes generally consistent with linguistic bias (LEB)	– Descriptions of W change more than those of M in all contexts – Descriptions of Asian/Black people change more than others (7/8, 1/8 contexts) – Little evidence of linguistic bias

6   ACCURACY VERSUS FAIRNESS – A TRADE-OFF?

We evaluated five proprietary image tagging algorithms' accuracy in using gender-appropriate tags, on a binary "male/female" spectrum; thus, our approach is similar in spirit to that of other recent studies (e.g., [15, 56, 77]). As the highly-controlled dataset we used, the Chicago Face Database, only features binary cisgender individuals, we could not consider how the five ITAs gendered non-binary and/or transgender individuals. The gender recognition accuracy for non-cisgender people (in particular, trans women, trans men, and non-binary individuals) was the focus of the recent work by Scheuerman and colleagues [93], who constructed a novel dataset of Instagram images with self-descriptions of gender for their analyses.

Going beyond an evaluation of accuracy, our analysis focused on understanding how *gender stereotyping* — based on the introduction of a background context suggesting a social role (e.g., caretaker, bride or groom) or occupation (e.g., mechanic, chef) — might affect the recognition of binary genders. The highly uniform CFD images were imposed upon eight different background contexts, and our question was whether this might aid or hinder gender recognition. Recognizing the ITAs as ML-based systems that have "hybrid human-machine behaviors" [89], as well as the strong human tendency to associate social contexts with gender, we expected that imposing a CFD image upon a background image that is stereotype-congruent, would help the ITA recognize the person's gender. However, we found that many ITAs could not consistently recognize both background and gender. In fact, only Amazon and Clarifai showed consistency here.

Thus, our analyses of the differences between descriptions of the baseline vs. contextual images, focused on Amazon and Clarifai. It is clear from our findings that with both of these services, the demographics of the people depicted (even in such controlled, standardized settings) affected the results. In particular, the treatment of images of women and non-white races (Clarifai & Amazon on images of Black people; Amazon on images of Asian people) were affected the most by a change in the background context. In other words, ITAs' descriptions of e.g., white men, are much more consistent across background contexts, than are those of other race/gender groups. This finding is in line with that of Scheuerman and colleagues, who found evidence that all four services they studied (Amazon, Clarifai, Watson, Microsoft) "understand" "male" as something more specific and bounded than "female" [93, p.15]. Likewise, in our analysis, the descriptions of images of white men in context, were more bounded (i.e., changed less with respect to the baseline) than those of other social groups.

The contexts that were strongly gendered according to our crowd study — wedding dress shop and mechanic's garage — proved most interesting. Here, it can be said that Clarifai's descriptions of people, more so than those of Amazon, reflected human gender stereotyping behavior. In particular, Clarifai's differential use of concepts, in that images of more stereotype-congruent people were described with more abstract and less concrete tags as compared to stereotype-incongruent people, were as predicted by the theory of Linguistic Expectancy Bias. In other words, the "communication style" used by Clarifai in describing these individuals, mimicked the proposed human tendency to be more inferential when describing stereotype-congruent people. It is also of note that the background contexts aided Clarifai's recognition of women in all four "feminine" backgrounds, while its recognition of men in the "masculine" backgrounds remained nearly the same or improved.

This highlights the difference between being "human-like" (behaving in a manner we have come to expect from human interactions) and being fair (behaving in a manner that is socially just) or at the very least, that respects *group fairness*, such that similar error rates are experienced across social groups [86]. Unfortunately, as seen from our results, this creates a trade-off between two behaviors in an algorithmic tagging service: shall it be "human-like" by formalizing gender stereotypes, and

increase its gender inference performance? Or shall it avoid reproducing stereotypes, regardless of how much this strategy lowers the accuracy of gender inferences?

Clarifai, as we have seen, is a good case study for one end of this spectrum. On the other hand, Google Cloud Vision[25] presents an equivalent alternative case. As mentioned in Section 4, our dataset was collected in December 2019. Originally, we included six ITAs, collecting tags provided by Google in addition to the other five services. Subsequently, we had to remove Google from our dataset, as in February 2020, it changed its service, removing all gender-related tags.[26] An email sent to users stated that the company made this decision based on the fact that "a person's gender cannot be inferred by appearance" (in line with e.g., [93]) and in order to align the company's practices with their principle to "avoid creating or reinforcing unfair bias."[27]

The accuracy versus fairness issue is one that is well-discussed in the machine learning community, particularly in the context of classification problems (e.g., [32, 39, 74]). Significant challenges for fair-ML include the fact that it is difficult — if not impossible — to satisfy multiple notions of fairness [63]. However, the matter is obviously much more complex for systems deployed into real-world, social contexts, which more often than not involve multiple stakeholders, and where the stakes are often high.

For instance, one domain where split-second decisions making use of gender stereotypes, technology, and fairness collide are body scanners at airport security [22]. The human security officers make use of their mental model of gender presentations to select one of two settings on the device controlling the body scanner, establishing some expectations for the body that is about to walk in. For many people, this is a regular and unremarkable part of the security process; and for those people the process works accurately and quickly. The vast majority of people passing through the system, who are not a security risk, are predicted as such. However, for people who do not fall into the expectations formalized into the body scanner (or "within an acceptable range of "deviance" from a normative binary body type" [22]) this process is very often inaccurate, stressful, and time-consuming.

In addition to this, people with multiple marginalized social identities, like people of color, Muslims, immigrants, and/or people with disabilities (who are simultaneously at the intersection of many oppressive systems) may have an even harder time due to the frequent and expected false positives. This is the double-edged sword Kleinberg et al. [63] discuss, as multiple constraints are near impossible to satisfy simultaneously. Ultimately, in the design of airport scanners, there was likely a choice between accuracy and fairness, and the creators most likely gave emphasis to accuracy; the false positives in the overall system are rare, and preferable to false negatives, so the system is deemed suitable and is deployed as such.

6.1 Limitations

Just as our auditing approach offers some improvements over previous techniques designed to evaluate the behaviors of algorithmic image tagging, it has its limitations as well. By creating our own artificial, composite images rather than studying images collected from the wild, we increase our control over the properties of the images analyzed. However, the created images may look somewhat artificial to a human analyst. In addition, in the present study we have used only one background image per context; future work should incorporate multiple images per context, in order to control for the fact that these contexts (e.g., baby nurseries) do not all look the same. This would also allow us to explore additional parameters of interest in image analysis, such as the

---

[25]cloud.google.com/products/ai/

[26]businessinsider.com/google-cloud-vision-api-wont-tag-images-by-gender-2020-2/

[27]ai.google/principles

potential impact of colors (such as blue or pink in a nursery) on the recognition of gender. It should also be noted that it was more difficult than anticipated to find royalty-free background images that were suitable for the task. Many of these images were most certainly taken in high-income environments (see both the wedding dress shop and the baby nursery images, in Figures 3 and 2), underscoring the need to expand to multiple images for a given social context.

Lastly, as with previous work, we have studied five specific proprietary image tagging algorithms, opting for their "general model" whenever available. Some services have other ITA models that are specific to a context (e.g., photos of food or fashion), as well as *facial recognition* algorithms (which are often trained specifically for gender analysis). These other services may perform differently, as compared to the "general models" we have studied herein.

7 CONCLUSION

The current research highlights the importance of not taking ML-based systems' output for granted. Although cognitive services provide a convenient means for developers to enhance their applications and services with state-of-the-art capabilities, there is a real potential to carry any inherent biases downstream, thus impacting other users. Likewise, researchers using proprietary ITAs to understand large-scale visual communication patterns on social platforms, need to be especially sensitive when processing images depicting people.

Through controlled experiments, we can attempt to understand and analyze the behavior of proprietary, black box systems, with respect to how they treat different social groups, and whether they do so in a manner that is consistent with human social norms. At the same time, it is important to appreciate that the behaviors observed during an audit may not translate fully to real world performance, where ITAs analyze images in the wild. Following research in psychology, stereotypes drive the way people interact with each other, the assumptions people make of others and in some occasions, people's future decisions for fulfilling society's stereotypical expectations. There is enough evidence in the literature that social stereotypes make their way into cognitive services and that they are able to at least partly affect users' perception. In addition, in extreme cases, ITAs can embody modern versions of physiognomy, a debunked Victorian pseudoscience that has been experiencing recent rebirth [20, 99]. As such, auditing ITAs, both in terms of their training data [88] as well as their output, ethically evaluates the role such services play.

We have seen that the cognitive services that were audited here, are not consistent in how they treat gender (men/women) with or without a social context. Among those that were more consistent, Clarifai "sees" more women and men when they appear in a stereotypically feminine or masculine context respectively, while Amazon "sees" more men in that context and understands context in a better way when gender is perceived. As a result of the controlled experimental approach adopted in this work, we were able to capture evidence of inconsistencies – violations of the group fairness notion – by the cognitive services that need further research. The findings based on race were particularly notable, as both Clarifai and Amazon's descriptions of images for images of Blacks and Asians were less stable, changing with the background context, as compared to others.

The recent approach of Google to remove gender-related tags from ITA tag-sets, while done to mitigate social bias, does not provide a comprehensive solution to the underlying problem (e.g., [10]). Service creators as well as developers who are using the services in their systems need to understand that there is no one-size-fits-all approach when comes to deciding on fairness versus accuracy. There are applications where gender inferences can be irrelevant and should be taken out, but there are also scenarios, like dating applications [54], where gender and/or race stereotyping can have a huge impact on someone's life and well-being. Therefore, auditing using multiple datasets is important for understanding the benefits and harm that a black box service might cause when deployed and real people are affected by its output.

ACKNOWLEDGMENTS

REFERENCES

[1] Vernon L Allen and Evert Van de Vliert. 1984. A role theoretical perspective on transitional processes. In *Role transitions*. Springer, 3–18.

[2] Tawfiq Ammari and Sarita Schoenebeck. 2016. "Thanks for your interest in our Facebook group, but it's only for dads" Social Roles of Stay-at-Home Dads. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1363–1375.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.

[4] Cory L Armstrong and Michelle R Nelson. 2005. How newspaper sources trigger gender stereotypes. *Journalism & Mass Communication Quarterly* 82, 4 (2005), 820–837.

[5] Saeideh Bakhshi, Lyndon Kennedy, Eric Gilbert, and David A Shamma. 2019. Filtered Food and Nofilter Landscapes in Online Photography: The Role of Content and Visual Effects in Photo Engagement. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 80–90.

[6] Andrea Ballatore, Michela Bertolotto, and David C Wilson. 2014. An evaluative baseline for geo-semantic relatedness and similarity. *GeoInformatica* 18, 4 (2014), 747–767.

[7] Pınar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2019. Social b (eye) as: Human and machine descriptions of people images. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 583–591.

[8] CJ Beukeboom, JP Forgas, O Vincze, and J Laszlo. 2014. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social Cognition and Communication* (2014), 313–330.

[9] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 514–524.

[10] Abeba Birhane and Fred Cummins. 2019. Algorithmic Injustices: Towards a Relational Ethics. *arXiv preprint arXiv:1912.07376* (2019).

[11] Philipp Blandfort, Desmond U Patton, William R Frey, Svebor Karaman, Surabhi Bhargava, Fei-Tzin Lee, Siddharth Varia, Chris Kedzie, Michael B Gaskell, Rossano Schifanella, et al. 2019. Multimodal social media analysis for gang violence prevention. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 114–124.

[12] Galen V Bodenhausen. 1993. Emotions, arousal, and stereotypic judgments: A heuristic model of affect and stereotyping. In *Affect, cognition and stereotyping*. Elsevier, 13–37.

[13] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[14] Hege Eggen Børve and Elin Børve. 2017. Rooms with gender: physical environment and play culture in kindergarten. *Early Child Development and Care* 187, 5-6 (2017), 1069–1081.

[15] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[16] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[17] Carrie J Cai and Philip J Guo. 2019. Software Developers Learning Machine Learning: Motivations, Hurdles, and Desires. In *2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 25–34.

[18] Mary Ann Cejka and Alice H Eagly. 1999. Gender-stereotypic images of occupations correspond to the sex segregation of employment. *Personality and social psychology bulletin* 25, 4 (1999), 413–423.

[19] Deborah Chavez. 1985. Perpetuation of gender inequality: A content analysis of comic strips. *Sex Roles* 13, 1-2 (1985), 93–102.

[20] Sahil Chinoy. 2019. The Racist History Behind Facial Recognition. *The New York Times [Internet]* 883 (2019).

[21] Scott Coltrane and Melinda Messineo. 2000. The perpetuation of subtle prejudice: Race and gender imagery in 1990s television advertising. *Sex roles* 42, 5-6 (2000), 363–389.

[22] Sasha Costanza-Chock. 2018. Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science* (16 7 2018). https://doi.org/10.21428/96c8d426 https://jods.mitpress.mit.edu/pub/costanza-chock.

[23] Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology* 40 (2008), 61–149.

[24] Amy JC Cuddy, Peter Glick, and Anna Beninger. 2011. The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in organizational behavior* 31 (2011), 73–98.

[25] David Danks and Alex John London. 2017. Algorithmic Bias in Autonomous Systems.. In *IJCAI*. 4691–4697.

[26] Abhijit Das, Antitza Dantcheva, and Francois Bremond. 2018. Mitigating Bias in Gender, Age and Ethnicity Classification: a Multi-Task Convolution Neural Network Approach. In *ECCVW 2018-European Conference of Computer Vision Workshops*.

[27] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 52–59.

[28] Julia Deeb-Swihart, Christopher Polack, Eric Gilbert, and Irfan Essa. 2017. Selfie-presentation in everyday life: A large-scale characterization of selfie contexts on instagram. In *Eleventh International AAAI Conference on Web and Social Media*.

[29] Alessandro Del Sole. 2018. Introducing microsoft cognitive services. In *Microsoft Computer Vision APIs Distilled*. Springer, 1–4.

[30] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62.

[31] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 278–288.

[32] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*. 119–133.

[33] Alice H Eagly and Antonio Mladinic. 1994. Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European review of social psychology* 5, 1 (1994), 1–35.

[34] Alice H Eagly and Wendy Wood. 1982. Inferred sex differences in status as a determinant of gender stereotypes about social influence. *Journal of personality and social psychology* 43, 5 (1982), 915.

[35] Alice H. Eagly and Wendy Wood. 2012. Social Role Theory. In *Handbook of theories of social psychology*, John C Turner and Katherine J Reynolds (Eds.). Vol. 2. Sage London, Chapter 49, 458–476. https://doi.org/10.4135/9781446249222.n49

[36] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; things can be worse than they appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Eleventh International AAAI Conference on Web and Social Media (ICWSM)*. 62–71.

[37] Klaus Fiedler and Gün R Semin. 1988. On the causal information conveyed by different interpersonal verbs: The role of implicit sentence context. *Social Cognition* 6, 1 (1988), 21–39.

[38] Katherine Fink. 2018. Opening the government's black boxes: freedom of information and algorithmic accountability. *Information, Communication & Society* 21, 10 (2018), 1453–1471.

[39] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 144–152.

[40] Susan T Fiske, Amy JC Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences* 11, 2 (2007), 77–83.

[41] Howard N Garb. 1994. Cognitive heuristics and biases in personality assessment. In *Applications of heuristics and biases to social issues*. Springer, 73–90.

[42] Venkata Rama Kiran Garimella, Abdulrahman Alfayad, and Ingmar Weber. 2016. Social media image analysis for public health. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5543–5547.

[43] CS Garrett, P Lynne Ein, and Leslie Tremaine. 1977. The development of gender stereotyping of adult occupations in elementary school children. *Child Development* (1977), 507–512.

[44] Patrick J Grother, Mei L Ngan, and Kayee K Hanaoka. 2019. NISTIR 8280 Face Recognition Vendor Test Part 3: Demographic Effects. *National Institute of Standards and Technology* (2019).

[45] Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. 2019. What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 236–246.

[46] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–13.

[47] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1914–1933.

[48] S Alexander Haslam, John C Turner, Penelope J Oakes, Katherine J Reynolds, and Bertjan Doosje. 2002. From personal pictures in the head to collective tools in the world: How shared stereotypes allow groups to represent and change social reality. (2002).

[49] Madeline E Heilman and Alice H Eagly. 2008. Gender stereotypes are alive, well, and busy producing workplace discrimination. *Industrial and Organizational Psychology* 1, 4 (2008), 393–398.

[50] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*. Springer, 793–811.

[51] Susan C Herring. 2009. Web content analysis: Expanding the paradigm. In *International handbook of Internet research*. Springer, 233–249.

[52] Marc Hooghe, Laura Jacobs, and Ellen Claes. 2015. Enduring gender bias in reporting on political elite positions: Media coverage of female MPs in Belgian news broadcasts (2003–2011). *The International Journal of Press/Politics* 20, 4 (2015), 395–414.

[53] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What we instagram: A first analysis of instagram photo content and user types. In *Eighth International AAAI conference on weblogs and social media*.

[54] Jevan A Hutson, Jessie G Taft, Solon Barocas, and Karen Levy. 2018. Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.

[55] Ji Yoon Jang, Sangyoon Lee, and Byungjoo Lee. 2019. Quantification of Gender Representation Bias in Commercial Films based on Image Analysis. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–29.

[56] 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 *Twelfth International AAAI Conference on Web and Social Media*.

[57] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3819–3828.

[58] John E Kelly. 2015. Computing, cognition and the future of knowing. *Whitepaper, IBM Reseach* 2 (2015).

[59] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[60] Kishor S Kinage and SG Bhirud. 2008. Racial inconsistency in face recognition. In *SPIT-IEEE Colloquium and International Conference*, Vol. 1. 78–81.

[61] Katherine N Kinnick. 1998. Gender bias in newspaper profiles of 1996 Olympic athletes: A content analysis of five major dailies. *Women's Studies in Communication* 21, 2 (1998), 212–237.

[62] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6 (2012), 1789–1801.

[63] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS), 2017*. http://arxiv.org/abs/1609.05807

[64] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the 2015 Internet Measurement Conference (IMC '15)*. ACM, New York, NY, USA, 121–127. https://doi.org/10.1145/2815675.2815714 event-place: Tokyo, Japan.

[65] Enes Kocabey, Ferda Ofli, Javier Marin, Antonio Torralba, and Ingmar Weber. 2018. Using computer vision to study the effects of bmi on online popularity and weight-based homophily. In *International Conference on Social Informatics*. Springer, 129–138.

[66] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 417–432. https://doi.org/10.1145/2998181.2998321 event-place: Portland, Oregon, USA.

[67] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 313–322.

[68] Martha M Lauzen, David M Dozier, and Nora Horan. 2008. Constructing gender stereotypes through social roles in prime-time television. *Journal of Broadcasting & Electronic Media* 52, 2 (2008), 200–214.

[69] Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 34–42.

[70] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E Moghaddam, and Lyle Ungar. 2016. Analyzing personality through social media profile picture choice. In *Tenth international AAAI conference on web and social media*.

[71] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47, 4 (2015), 1122–1135.

[72] Anne Maass, Daniela Salvi, Luciano Arcuri, and Gün R Semin. 1989. Language use in intergroup contexts: The linguistic intergroup bias. *Journal of personality and social psychology* 57, 6 (1989), 981.

[73] Gabriel Magno, Camila Souza Araujo, Wagner Meira Jr., and Virgilio Almeida. 2016. Stereotypes in Search Engine Results: Understanding The Role of Local and Global Factors. *arXiv:1609.05413 [cs]* (Sept. 2016). http://arxiv.org/abs/1609.05413 arXiv: 1609.05413.

[74] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*. 107–118.

[75] Debra Merskin. 2007. Three faces of Eva: Perpetuation of the hot-Latina stereotype in Desperate Housewives. *The Howard Journal of Communications* 18, 2 (2007), 133–151.

[76] Abbe Mowshowitz and Akira Kawaguchi. 2005. Measuring Search Engine Bias. *Inf. Process. Manage.* 41, 5 (Sept. 2005), 1193–1205. https://doi.org/10.1016/j.ipm.2004.05.005

[77] 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. *arXiv preprint arXiv:1812.00099* (2018).

[78] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*.

[79] Kalia Orphanou, Jahna Otterbacher, Styliani Kleanthous, Khuyagbaatar Batsuren, Fausto Giunchiglia, Veronika Bogina, Avital Shulner Tal, Alan Hartman, and Tsvi Kuflik. 2020. Deliverable D3.4 - Survey Article. http://www.cycat.io/wp-content/uploads/2020/06/D3.4_Survey_Article_NV.pdf Project deliverable, H2020 CyCAT (810105).

[80] Jahna Otterbacher. 2015. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1955–1964.

[81] Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. 2019. How Do We Talk about Other People? Group (Un) Fairness in Natural Language Image Descriptions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 106–114.

[82] J. Otterbacher, J. Bates, and P. D. Clough. 2017. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3025453.3025727

[83] Neelamdhab Padhy, RP Singh, and Suresh Chandra Satapathy. 2018. Software reusability metrics estimation: algorithms, models and optimization techniques. *Computers & Electrical Engineering* 69 (2018), 653–668.

[84] Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206* (2017).

[85] Kayur Patel, James Fogarty, James A Landay, and Beverly Harrison. 2008. Investigating statistical machine learning as a tool for software development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 667–676.

[86] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM, 581–592.

[87] David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2011).

[88] Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? arXiv:2006.16923 [cs.CY]

[89] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477.

[90] Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 74–79.

[91] Mohamed Aymen Saied, Ali Ouni, Houari Sahraoui, Raula Gaikovina Kula, Katsuro Inoue, and David Lo. 2018. Improving reusability of software libraries through usage pattern mining. *Journal of Systems and Software* 145 (2018), 164–179.

[92] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* (2014), 1–23.

[93] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[94] Christine K Shenouda and Judith H Danovitch. 2014. Effects of gender stereotypes and stereotype threat on children's performance on a spatial task. *Revue internationale de psychologie sociale* 27, 3 (2014), 53–77.

[95] Amit Sheth, Hong Yung Yip, Arun Iyengar, and Paul Tepper. 2019. Cognitive services and intelligent chatbots: current perspectives and special issue introduction. *IEEE Internet Computing* 23, 2 (2019), 6–12.

[96] Eva H Shinar. 1975. Sexual stereotypes of occupations. *Journal of vocational behavior* 7, 1 (1975), 99–111.

[97] Vivek Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 2019. Female Librarians and Male Computer Programmers? Gender Bias in Occupational Images on Digital Media Platforms. *arXiv preprint arXiv:1912.05474* (2019).

[98] Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.

[99] Luke Stark. 2018. Facial recognition, emotion and race in animated social media. *First Monday* (2018).

[100] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Queue* 11, 3 (March 2013), 10:10–10:29. https://doi.org/10.1145/2460276.2460278

[101] Emiel Van Miltenburg. 2016. Stereotyping and Bias in the Flickr30K Dataset. In *Proceedings of the Workshop on Multimodal Corpora (MMC-2016)*. 1–4.

[102] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[103] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*.

[104] Michael J White and Gwendolen B White. 2006. Implicit and explicit occupational gender stereotypes. *Sex roles* 55, 3-4 (2006), 259–266.

[105] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097* (2019).

[106] Michele Wilson. 2017. Algorithms (and the) everyday. *Information, Communication & Society* 20, 1 (2017), 137–150.

[107] Piotr Winkielman, Jamin Halberstadt, Tedra Fazendeiro, and Steve Catty. 2006. Prototypes are attractive because they are easy on the mind. *Psychological science* 17, 9 (2006), 799–806.

[108] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

[109] Liqi Zhu and Gerd Gigerenzer. 2006. Children can solve Bayesian problems: The role of representation in mental computation. *Cognition* 98, 3 (2006), 287–308.