Person, Human, Neither: The Dehumanization Potential of Automated Image Tagging

Pınar Barlas
p.barlas@cyens.org.cy
CYENS Centre of Excellence
Nicosia, Cyprus

Kyriakos Kyriakou
k.kyriakou@cyens.org.cy
CYENS Centre of Excellence
Nicosia, Cyprus

Styliani Kleanthous*
styliani.kleanthous@ouc.ac.cy
Cyprus Center for Algorithmic Transparency
Open University of Cyprus
Nicosia, Cyprus

Jahna Otterbacher*
jahna.otterbacher@ouc.ac.cy
Cyprus Center for Algorithmic Transparency
Open University of Cyprus
Nicosia, Cyprus

ABSTRACT

Following the literature on dehumanization via technology, we audit six proprietary image tagging algorithms (ITAs) for their potential to perpetuate dehumanization. We examine the ITAs' outputs on a controlled dataset of images depicting a diverse group of people for tags that indicate the presence of a human in the image. Through an analysis of the (mis)use of these tags, we find that there are some individuals whose 'humanness' is not recognized by an ITA, and that these individuals are often from marginalized social groups. Finally, we compare these findings with the use of the 'face' tag, which can be used for surveillance, revealing that people's faces are often recognized by an ITA even when their 'humanness' is not. Overall, we highlight the subtle ways in which ITAs may inflict widespread, disparate harm, and emphasize the importance of considering the social context of the resulting application.

CCS CONCEPTS

• **Social and professional topics** → *User characteristics*; • **Human-centered computing**;

KEYWORDS

image tagging algorithms; dehumanization; science, technology, and society studies; critical computing

ACM Reference Format:
Pınar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2021. Person, Human, Neither: The Dehumanization Potential of Automated Image Tagging. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462567

> *"Humanity is a 'thing,' and [the oppressors] possess it as an exclusive right, as inherited property."*
> – Paulo Freire

1 INTRODUCTION

For many developers and small start-ups, training a machine learning model from scratch is difficult, expensive, and time-consuming. However, cognitively inspired algorithms have become commercialized (often called *Cognitive Services*), offering economical, easy-to-use solutions and creating what is being called the "Algorithm Economy."[1] Developers are no longer limited by time, money, skill, or equipment, and increasingly use complex algorithms developed by other companies to power their own applications.

In part due to the reputation of the providers, some people place more trust in algorithms than in conventional institutions [31]. Fueled also by the faith in "statistical objectivity" of machine learning (ML) models [17], developers and other end-users often take the outputs of such systems at face value. However, ML models that are trained on historical data can pick up on, embed, and reinforce existing structural inequality. For example, hiring algorithms are hailed as "unbiased" filtering systems focused only on features relevant to the job,[2] even though it has been shown that historical data[3] or the choice of candidate features that predict job performance (i.e. proxies) [34] will replicate existing harmful patterns.

As opposed to traditional software "breakdowns," errors in predictive ML applications do not come with a notification of abnormal functionality. Instead, they appear as regular outputs, and when left unchecked can escalate into widespread harm down the pipeline of an application. Researchers, upon seeing examples of harm in deployed applications and recognizing the potential for harm in others, have been taking a critical look. In systems that are opaque (e.g. with proprietary code, features that are not (publicly) available, or technical complexity impeding explainability), audits have been conducted by examining the outputs for a set of known inputs [44]. We find this method appropriate for our research and conduct an audit, checking for "group fairness" (i.e. disparate impact on social groups) [15] to examine the treatment of people by proprietary ML

*Also with CYENS Centre of Excellence, Cyprus.

AIES '21, May 19–21, 2021, Virtual Event, USA

[1] https://www.gartner.com/smarterwithgartner/the-algorithm-economy-will-start-a-huge-wave-of-innovation/
[2] https://www.hirevue.com/blog/hiring/hirevue-assessments-and-preventing-algorithmic-bias
[3] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

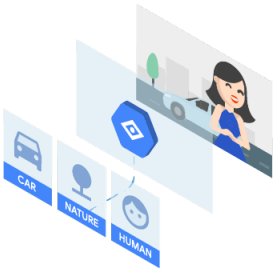

Figure 1: Illustration of the Google Vision API service, labeling the image of a person's face with the 'human' tag, found on: https://cloud.google.com/vision/; © Google.

services, particularly those which can be purchased as pre-trained models.

Many Cognitive Services focus on *image analysis*. As the number of images online grows, so does the need for automated processing and handling of those images. Humans are depicted in a great number of those images, and are often relevant to the end-goal of the processing. Therefore, many image analysis services focus specifically on the human face: recognizing and locating, inferring characteristics from, and matching with a previously-seen face. The use of facial detection/analysis/recognition services is growing, and with it, the numbers of audits revealing disparate error rates,[4] of inferences fueling harmful beliefs about already-marginalized communities [23, 50], and of large-scale use cases infringing on people's privacy and feelings of safety.[5,6]

We believe this has led to some neglect of the potentially harmful treatment of humans by other, more general, image analysis services. A popular class of Cognitive Service offering is image tagging algorithms (ITAs), in which the inferred content of an image is returned as a set of descriptive tags; one such service has the explanatory diagram in Figure 1, displayed prominently on the provider's website. In addition to reported uses (e.g. managing personal photo collections[7] and cultural heritage archives,[8] even applications for emotional[9] and physical health[10]), ITAs may be used in the pipeline of an application where the outputs are fed into yet another algorithm. When such an application will handle mixed image inputs (i.e. images that may or may not include humans), but will make decisions based on whether there is a human in the image, it is necessary to first inquire if any humans are detected in the image. The pre-trained ITA models often handle specialized topics, e.g. food or apparel.[11] However, we focus our audit on the "general models" of the ITAs, which are advertised to be a comprehensive solution recognizing many concepts and are most likely to be used to handle mixed image inputs in order to detect humans - even if they are not prominently depicted.

As with many other ML applications, ITAs have also been shown to embody social biases [e.g. 11, 25]. In two striking examples reported only a few months apart in 2015, first Flickr,[12] and then the newly-launched Google Photos application,[13] associated images of Black people with the tags "ape," "gorillas," and "animal," reflecting and reinforcing a centuries-old racist trope.[14] This trope, where Black people are likened to nonhuman primates, is a clear example of *dehumanization*. Dehumanization refers to the psychological concept of perceiving people as nonhuman, which can be a predictor for and enabler of violence [24, 42].

Given the trust people have in algorithms, the fact that ITAs are assigned the role of declaring what is in the image, and the harmful effects of machine learning models left unchecked, we audit six proprietary ITAs for their potential to dehumanize. We ask three research questions: one to discover how the ITAs interpret 'humanness,' one to explore whether this interpretation holds true for images of different social groups, and another to judge whether the facial detection of the depicted individuals took priority over their 'humanness' in the training of the ITAs.

2 BACKGROUND

We first review the foundational dehumanization literature in social and personality psychology, establishing an understanding of how dehumanization can appear in human-human interactions. Next, we examine the way social biases can and have become embedded in machine learning models, in addition to other harmful effects of technology. Finally, we merge the two perspectives, discussing work that draws connections between technology and dehumanization.

2.1 Dehumanization + humans

In social and personality psychology, dehumanization is described as one person (i.e. the perceiver) not seeing another person (i.e. the target) as fully human [22], or when the perceiver's behavior undermines basic elements of the target's personhood, such as identity and status [6]. There are different ways such perception and behavior can be observed.

2.1.1 Methods of dehumanization. Metaphor-based dehumanization is when the perceiver directly declares that the target *is nonhuman*, e.g. "that person is" an animal, or pest. Attribute-based dehumanization on the other hand, is when the perceiver declares that the target *has some nonhuman qualities*, e.g. "that person is" unthinking, or primal [28]. We choose to look at the metaphor-based dehumanization by examining the tags from Image Tagging Algorithms (ITAs) relating directly to 'humanness'; future work may look into the use of tags that describe (non)human qualities.

2.1.2 Levels of dehumanization. A sliding scale of dehumanizing behavior is discussed in the literature, ranging from 'everyday' or subtle dehumanization (e.g. physicians viewing patients as lacking individuation and agency [19]), all the way to 'severe' or blatant

[4] https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28
[5] https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html
[6] https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html
[7] https://www.skyfish.com/features/tagging
[8] https://customers.microsoft.com/en-in/story/nava-civilian-government-azure-services-hungary
[9] https://www.clarifai.com/blog/clarifai-featured-hack-improve-your-emotional-health-with-feelybot
[10] https://www.clarifai.com/blog/clarifai-featured-hack-how-healthy-or-unhealthy-is-your-meal-foodifai-knows
[11] https://www.clarifai.com/developers/pre-trained-models
[12] https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos
[13] https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas
[14] https://www.nytimes.com/2018/06/17/opinion/roseanne-racism-blacks-apes.html

dehumanization (e.g. calling a social group 'vermin') [24]. In the current work, we focus primarily on blatant dehumanization through explicit (mis)use of 'humanness' tags. However, we also look briefly at a related tag ('face'), the use of which can enable more subtle dehumanization. It must be noted that attribute- and metaphor-based dehumanization are interlinked, as are subtle and blatant dehumanization; all types of dehumanization discussed can lead to the same negative effects with varying magnitudes [24, 28].

2.1.3 Types of dehumanization. Researchers discussing dehumanization differentiate between *the type of nonhuman* to which the target is likened. One example is animalistic dehumanization, when the target is likened to animals [21]. Through animalistic dehumanization, the target is denied qualities that are 'unique to humans' such as civility, complex emotions, and intelligence; they are seen as uncultured, coarse, amoral, and/or childlike [21].

The other type is mechanistic dehumanization, where the target is likened to machines or objects [21]. In mechanistic dehumanization, the target is denied qualities of 'human nature'; such qualities are inborn, "core properties" of humans that are "prevalent within populations and universal across cultures" [20]. Attributes of 'human nature' include warmth, vitality, and depth, and as a result of their denial, the target is seen as interchangeable, lacking agency, and superficial [21]. As opposed to the Google Photos tagger described earlier which used animal tags on humans, we look at the misuse of tags that indicate 'humanness'; therefore, we are looking at the potential denial of 'humanity' – innate, 'core' properties of humans – to some individuals and not others, reflecting mechanistic dehumanization.

In addition, mechanistic dehumanization is often linked with objectification, which is when the target is seen only as a(n interchangeable) means to an end, or as nothing more than their body [29, 33, 42]. This is in line with the later part of our investigation, in which we examine whether biometric information may be 'prioritized' in the ITAs' outputs such that surveillance is valued over people's 'humanity'.

2.1.4 Effects of dehumanization. Researchers caution that "categorical denial of membership [of the target] in this most basic of superordinate identities – 'human' – signals otherness in a profound way that can have dire consequences" [24]. Indeed, it has been observed that dehumanization is linked to attitudes ranging from disregard and indifference [20] to discrimination, aggression, and violence [18, 20, 24, 42] on the part of the perceiver. The targets of dehumanization experience, in addition to the effects of the previously mentioned actions, "cognitive deconstructive states and feelings of sadness and anger" [6]. In Sections 2.2 and 2.3, we discuss the potential impact of the (mis)use of 'humanness' tags, which can enable these effects to be replicated and amplified through widely-used technology. In Section 5, we discuss the potential impact of specific 'humanness' tags we find in our study of six ITAs.

2.1.5 Disparate dehumanization. While any person can dehumanize another in interpersonal contexts, it has been found that 'humanness' characteristics are attributed to in- and out-groups differently [20], and that social groups with power and/or status are often seen to have such characteristics more than other groups [21]. In addition, dehumanization is already prevalent in certain domains

Figure 2: LF-200 and BM-026, two images from the CFD.

(women as sexual objects [42], inmates as numbers and athletes as statistics [24]) and it is sometimes said to be endemic to a domain because of its structure (e.g. inherent features of medical settings, relationships, and/or practices [19]). Therefore, in our analysis we question whether women and non-White (particularly, Black) individuals are more likely to be dehumanized through the ITAs' outputs, since this technology is created and used in an existing social structure that marginalizes said groups.

2.2 Humans + technology

Machine learning (ML) applications can fail at tasks that humans find easy. However, researchers have noticed that these errors are influenced by social hierarchies; matching a person's face in two different photos may fail more often for Black individuals[15] or detecting objects within an image may produce unsuccessful outputs more often for lower-income households [11]. Therefore, it's imperative that ML applications are examined to ensure the outputs do not reflect or reinforce social biases.

One challenge, as mentioned in the Introduction, is that ML applications are often 'black boxes' that prohibit third parties – or sometimes even the team creating the system – from understanding how data is processed and what calculations are made. A method proposed by Sandvig et al. to audit these systems 'from the outside' (i.e. by manipulating a set of inputs and analyzing the outputs) can give researchers an approximation of the process within an opaque system; many researchers have combined this with "group fairness" (i.e. disparate impact on social groups) [15] measures to determine whether outputs from proprietary systems exhibit social biases. We find these methods appropriate for our research and conduct an audit of six ITAs, using standardized images of diverse set of people to examine whether outputs of the ITAs differ depending on the social group (particularly race and gender) of the person depicted.

One common source of unwanted social biases in algorithmic systems is the use of a 'biased' dataset in training ML models: i.e. a dataset in which some qualities are over-/under- represented. In cases where historical data is used, existing structural inequality may be modeled; otherwise, it may be that the creators of the dataset selected a non-representative sample of the world [4, 10, 36]. In computer vision systems using images of humans, this could result from having more images of some social groups than others during training – which may result in a system that 'predicts' more accurately for that social group (e.g. white men having the highest accuracy for gender predictions in facial analysis [9]). To minimize the possibility of some of the input images resembling the training dataset more than others, we use tags collected using a controlled

[15] https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

dataset of people images, which represent diverse people in the same manner (Figure 2). Therefore, any differences in outputs will be more likely to result from the particularities of each individual depicted, rather than image composition. Specifically in annotated datasets, there are often systematic differences in the annotations for different social groups; crowdworkers annotating image datasets describe the same aspect of a human in many different ways [3, 35, 48], sometimes making inferences that reinforce a stereotype about the person's social group [49]. Consequently, we can expect patterns of human behavior such as dehumanization to appear in the outputs of the ITAs. In addition, since dehumanization may appear more often for marginalized communities, we expect a gender and/or race difference in error rates between images.

In addition, two 'biases' relating to human interpretation and use are relevant for understanding the dehumanization potential of ITAs. The first is the "transfer context bias" [10], where an algorithm is used in a context for which it is not intended. ITAs, which are inexpensive, 'out-of-the box' solutions available for anyone to plug into their application, are particularly vulnerable to this bias. The 'general models' of ITAs, which we examine, are especially marketed as solutions for developers regardless of the purpose or context of the final system. Given that it is impossible for the providers of the ITA to know every single potential use case, it is safe to assume many (if not all) use cases of 'general models' are vulnerable to biases resulting from a mismatch of the original training goals and the final use context of the algorithm. Secondly, "interpretation bias" refers to the unintended effects resulting from a misunderstanding (and therefore misuse) of the outputs by the end-user [10]. Many predictive computer vision applications, including ITAs, present their results with a success probability (often called a 'confidence' score) instead of definite predictions. Some ITAs automatically filter outputs based on a minimum threshold to ensure good results; others ask the user to do this. However, users can easily ignore this advice,[16] misinterpreting the existence of an output as the "correct" result and making further decisions based on this information. Therefore in our research we opt to ignore the scores associated with the 'humanness' tags, to simulate a developer who would be using such tags to automate decisions in the system.

It would be easy to think that, with so much evidence showing social biases can appear in and be replicated by machine learning applications, developers and the general public are hesitant to use such systems. However, the uncritical trust in algorithms – or at least the willingness to use algorithms to reach one's goals regardless of the error rate – is growing. In addition to some people showing a preference for algorithmic management than conventional institutions [31], it has been found that people are more likely to adhere to advice they think was given by an algorithm (as compared to that given by a person) [27]. This phenomenon, called *algorithmic authority*, is the tendency for people to "regard as authoritative an unmanaged process of extracting value from diverse, untrustworthy sources" [47]. In other words, algorithmic authority refers to the power of algorithms to present potentially untrustworthy data (or calculations) in a different way, leading some end users to believe that it is "the truth." An example often discussed is the Google Search ranking algorithm, which automatically legitimizes some sources of information over others, inadvertently impacting what the end user will believe about the topic searched [30]. In the case of ITAs, this translates into uncritical acceptance and use of tags received for a set of images by the end-users.

This behavior is partly fueled by the belief that algorithms are objective, unbiased tools that perform calculations insulated from the social and political context of the systems of which they are a part. However, all measurements (and therefore data and calculations based on those measurements) embody values which prescribe what gets measured and how [39]. For example, in the first years of color photography, it was believed that cameras produced unbiased representations of the world. However, in photographs taken using these cameras, people with dark skin would be rendered invisible, as the process had been calibrated using images of people with light skin [41]. Goldenfein argues that the "previous claims about the 'mechanical objectivity' of the camera have been displaced into the computational objectivity of statistics" through computer vision applications using ML, where neutrality can be used (consciously or not) to "launder" ideologies that may be harmful [17].

One of the most famous – and most harmful – examples of this phenomenon is physiognomy, the 'field' that claims a person's abstract characteristics can be interpreted from their appearance (often, their face). Goldenfein points to one of the first experiments connected to computer vision, in which Galton manually superimposed the portrait photographs of different criminals, hoping to end up with the prototypical criminal face [16]. Almost 140 years later, researchers trained a machine learning model on images of criminals to try and detect 'criminality' in faces [52]. Other types of computer vision applications have emerged that analyze people's images (often their face) to infer an abstract quality about them – like personality [53], gender [23], sexual orientation [50], or emotional state [5] – even if it has not been tested whether a connection between the abstract quality and the stimulus exists [17]. This is in line with the goal of many ML systems, which use "dynamic, pattern-based abstractions" to find "actionable indices of 'who we are'" [8]. Therefore, it is important to examine whether and how computer vision systems, such as the ITAs, make claims about people's 'humanness' and the impact of such claims being made through automated image analysis systems.

Despite their premise sometimes being unproven, and their potential dangers explicitly researched, such systems are already implemented and affect social groups differently. As Birhane and Cummins state, "algorithmic classifications and predictions give an advantage or they give suffering" [8]. People who have experienced the negative impacts of algorithmic treatment – and know about it – say "wrong labels, misclassification, and misinterpretation of data profiles are life-changing events" [38]. In examining image tagging algorithms, we are looking at a type of software which can be used in a wide range of applications from autonomous vehicle

[16] https://gizmodo.com/defense-of-amazons-face-recognition-tool-undermined-by-1832238149

parking[17] to personal/social functions.[18,19] In the former context, failure to recognize a human can lead to physically dangerous consequences, with some research already showing such errors may appear more often for darker-skinned individuals [51]. In the latter context, failure to recognize a human in an image (exacerbated by the presence of other tags, such as 'gorilla,' 'ape,' and 'animal') can, and has, affected the social perception of the person in the image, reinforcing a racist dehumanizing stereotype.

It is our hope that this public, named audit [like those before ours, e.g. 40] will motivate service providers to not only improve their performance on human recognition, but also to critically think about the broader dehumanizing uses and effects of their – existing *and* future – applications. To examine the more indirect types of dehumanization, we question what it means for an ITA to even have a tag that declares the existence of a 'human' in the image. As hinted at earlier in this section, people creating datasets/systems have the power to decide what kind of information is valued and even what is 'knowable' about the people represented by the data [13, 17], embedding their viewpoints and ideologies into the systems in various manners. Knowing also that measurements – and databases built with those measurements – determine what actions can be taken, and from there shape the way we see the world, we question whether as it "[became] normal to think of trees as lumber" [39], if it may also become normal to think of humans in 'datafied' ways.

2.3 Dehumanization + technology

Dehumanization and technology have been discussed together in various ways. Some research positions dehumanization as the counterpart to anthropomorphism, focusing on how humans may be the 'perceiver,' and technology the 'target,' of a sort of *reverse*-dehumanization [43, 46]. However, more relevant to our purposes is the literature which questions whether technology may take on the role of the dehumanizing 'perceiver,' while the humans represented (or replaced) by the data may be 'targets' of dehumanization.

Beliefs about the dehumanizing effects of technology were listed as one aspect of 'computer anxiety,' where questions related to technology causing isolation or destroying creativity [7]. Similarly, it has been observed by social and personality psychologists that dehumanizing implications of technology often discuss the denial of 'human nature' attributes to people [20], which follows the definition of mechanistic dehumanization.

Often, when the word dehumanization appears in a technology-related work, it refers to automation resulting in the removal of a human from the process, i.e. de-humanization [2, 13, 37]. Even when there is a connection between de-humanization of a process and the resulting dehumanizing effect, this connection may only be implicit. For example, Allwood discusses the digitalization (dehumanization) of services involving human contact, where people like cashiers, bank tellers, and customer service representatives are replaced with technology performing the same task [1]. While the author never explicitly names the implication as dehumanization, it is clear that the people in these roles are reduced to their transactional function, and implied to bring nothing else to the position than performing a series of actions. This reduction of people to a 'means to an end' exemplifies mechanistic dehumanization/objectification as discussed in Section 2.1. The mechanistic dehumanization through de-humanization phenomenon is discussed explicitly by Ekbia and Nardi, who describe how crowdworkers are 'rendered as bits of algorithmic function,' 'forcing the labor relation [further] along the path of ruthless objectification' [14].

The works which are explicit about the connection between de-humanization and dehumanization, however, often discuss the automation of human judgment in a critical decision-making process [26, 38]. Similarly, some work in areas such as medicine and autonomous weapons talk about how technology can create psychological and moral distance between the perceiver and the target, facilitating dehumanization through moral disengagement [2, 19]. Computer vision is increasingly part of dual-use applications,[20] so it is not a big leap to imagine the possibility that features of image tagging algorithms may also be used to inflict harm.[21,22]

More relevant to image tagging algorithms, however, are those cases that discuss dehumanizing effects of technology - whether or not it is referred to as such. In fact, we find that dehumanization via technology is far more common than the appearance of the word in technical literature. It is explicitly discussed that, when a system makes decisions based on data points instead of humans, the system is perceived as dehumanizing [26, 38]. Along with the moral disengagement effect of technology, the treatment of humans as data points can enable/automate large-scale systematic maltreatment of people, particularly marginalized social groups.[23,24]

Still, many examples can be found of people feeling like a statistic, number, data point, or "dataset of a future dataset" [38]. This is certainly not surprising as, at times, systems use scores to evaluate people "much like shoppers rate products on Amazon."[25] Researchers, referring to this pattern of representing people as scores or proxies, discuss how such numerical values 'inadequately capture or represent individuals,' [17] often resulting in distortion of the person represented (or now replaced) by the data [34, 38]. Being reduced to a functionality or "[node] of information production" [12] exemplifies the definition of mechanistic dehumanization/objectification: "the abstraction of [one's] experience and full humanity into categories, types, and ratings is a form of dehumanization" [38].

The complexity of humans, the contexts that algorithms are used in, and the interplay of these factors make it impossible to declare any one way of algorithmic calculation (or one manner of using 'humanness' tags, in the case of ITAs) is 'fair' in all cases. Just as the misrecognition of people may prevent them from receiving a

[17] https://www.clarifai.com/blog/clarifai-featured-hack-val.ai-is-a-parking-app-for-your-self-driving-car
[18] https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos
[19] www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas
[20] https://www.aclu.org/blog/national-security/targeted-killing/i-quit-my-job-protest-my-companys-work-building-killer
[21] https://theintercept.com/2018/09/06/nypd-surveillance-camera-skin-tone-search/
[22] https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html
[23] https://gizmodo.com/how-ibm-technology-jump-started-the-holocaust-5812025
[24] https://www.theguardian.com/us-news/2019/jul/11/amazon-ice-protest-immigrant-tech
[25] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

benefit of an algorithm, in some harmful use cases such as (disproportionate) surveillance,[26] errors and misrecognition may end up inadvertently benefiting marginalized communities.[27] Biometric systems, including and especially those used for surveillance, often use faces in images to automate calculations and decisions. Therefore, in order to question whether the ITAs may enable treatment of people in dehumanizing ways, we compare our findings of the 'humanness' tags to those indicating the presence of a 'face' in the image. This 'datafication' of humans is not a problem that can be solved easily – and not with (more) technical 'solutions' or audits.[28] That is why we refuse to simply focus on accuracy and instead investigate "why we find the patterns that we do" [8], looking at how 'humanness' and 'face' may be encoded in the ITAs through their use of such tags. We also invite creators of image analysis algorithms to question what types of applications they are enabling, and which ones they are restricting, through the outputs they make available.

2.4 Research questions

With this knowledge, in order to examine the potential for image tagging algorithms (ITAs) to dehumanize, we ask the following research questions:

(1) What vocabulary do the image tagging algorithms use to identify humans ('humanness tags')? What kind of dehumanizing effect may an error have?
(2) Do the ITAs apply this vocabulary to every photo in a controlled, diverse dataset of people images? Do errors appear for any social groups more than others?
(3) Do the ITAs identify faces? If so, are there any photos where the face is identified but the 'human' isn't?

3 METHODOLOGY

To answer our research questions we use the tags in the Social B(eye)as Dataset (SBD) [3], collected from six proprietary ITAs for the images in the Chicago Face Database [32], a diverse set of controlled portrait photographs. In this section, we briefly describe our methodology for the analysis, although more information for the input images and the tag collection can be found in the original papers for the respective datasets.

3.1 Input images

To collect the tags for the SBD, we had used a dataset of 597 controlled portrait photographs, depicting diverse individuals in the same manner in front of a white background and with neutral facial expressions, from the Chicago Face Database (CFD).[29] The CFD was created by Ma et al. for use in scientific research, particularly in psychology, where the researchers saw a need for a standardized, demographically-diverse dataset of faces [32]. The controlled manner in which the people are represented, along with demographic information about the individuals depicted, help compare computer vision algorithms' treatment of social groups in a systematic manner. An example from the CFD can be seen in Figure 2 in an earlier section, while the self-reported race and gender of the people depicted in the images the images (four and two mutually exclusive categories respectively) can be found in Table 1.

Table 1: The race and gender of the people depicted in the input images from the CFD.

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

3.2 Image tagging algorithms

We evaluate the outputs we had collected for these images from the following proprietary and opaque ITAs in 2018 [3]:

- Amazon Rekognition Image[30] (hereon: Amazon, A)
- Clarifai[31] (hereon: Clarifai, C)
- Google Cloud Vision[32] (hereon: Google, G)
- Imagga Auto-tagging[33] (hereon: Imagga, I)
- Microsoft Computer Vision[34] (hereon: Microsoft, M)
- IBM Watson Visual Recognition[35] (hereon: Watson, W)

In all six cases, we had used the 'general model' even when other models were offered, and we opt to use the tags without their respective confidence scores. Given the realistic uses of ITAs (Section 2.2), this data suits our needs for this analysis. However, Google's use of tags relevant to our research have been updated since their creation of this dataset; please see Section 5.1 for details. It is important for RQ3 to also note that all six providers also offer facial detection, analysis, and/or recognition/comparison services, as a separate application from the ITAs that we audit.

3.3 Analysis methods

3.3.1 RQ1. To find the vocabulary of the ITAs to identify humans, which we call 'humanness tags,' we gathered the tags which indicated the presence of a person. While there is a typology of tags accompanying the SBD [3], there was no cluster that suited our needs. Therefore, given the relatively small number of total tags, we opted for a manual analysis. Three researchers each conducted an analysis of the unique tags separately, identifying tags that directly indicated the presence of a human in the image. The researchers compared their results afterwards, opting to remove from the analysis any tags that signified an additional concept (e.g. 'woman' (gender, age), 'person with mohawk' (hairstyle)). For ITAs with more than one distinct 'humanness' tag, we conducted a co-occurrence analysis to discover relationships between the use of such tags.

[26] http://gutsmagazine.ca/watched-and-not-seen/
[27] https://www.thedailybeast.com/bots-are-terrible-at-recognizing-black-faces-lets-keep-it-that-way
[28] https://www.vox.com/future-perfect/2019/4/19/18412674/ai-bias-facial-recognition-black-gay-transgender
[29] https://chicagofaces.org/default/
[30] https://aws.amazon.com/rekognition/image-features/
[31] https://clarifai.com/developer/guide/
[32] https://cloud.google.com/products/ai/
[33] https://imagga.com/solutions/auto-tagging.html
[34] https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/
[35] https://www.ibm.com/watson/services/visual-recognition/

3.3.2 RQ2. To examine whether there were images of individuals which did not receive 'humanness' tags, we count the number of images that received the 'humanness' tags (i.e. frequency), comparing it to the full set of images. In addition, we inspect the demographics of the people depicted in the images which lack 'humanness' tags, to discover whether such errors are concentrated on any social groups in particular.

It is important to note here that we do not calculate overall percentages of error or run statistical tests of differences. The dehumanizing real-world effects of the ITAs' behavior make calculations of proportions irrelevant. In fact, movements/companies pushing for digitalization often favor discussions of 'high accuracy' of their systems and keep quiet on (the number of) people affected negatively by such automation [1]. So, even if an ITA has a 'low' error rate, any non-zero findings cannot be ignored. We will discuss this further in Section 5.

3.3.3 RQ3. Finally, to question whether dehumanizing treatment of people may be facilitated over a more 'humanized' approach, we compare the use of 'humanness' tags to tags indicating a 'face' in the image, focusing on the number of images in which a face may be identified while 'humanness' is not.

4 FINDINGS

4.1 RQ1: Calculating and tagging 'humanness'

We find that all six ITAs have at least one tag to identify a human in the image ('humanness tags'), as seen in Table 2. The most common tag was 'person,' used by five ITAs with the notable exception of Clarifai, which instead used 'no person.' Four ITAs also used 'people,' although our images only depicted one individual.

Two ITAs, Amazon and Google, used the 'human' tag in addition to the 'person' tag. While Amazon used both tags on every image in the dataset, Google used the tags on a small portion of the images. A co-occurrence analysis revealed that no image received both tags, implying that the Google ITA's 'calculations' of the human and person tags are distinct. We also found Watson had a tag ('people (face)') that was textually a combination of two other tags of interest. Our co-occurrence analysis on these three tags (Figure 3) showed no pattern to indicate one tag is the subset of another. Therefore, we can conclude that the three tags are made up of distinct calculations in the ITA's model.

Table 2: "Humanness tags" and frequencies by each ITA. *Includes the two images with the 'no person' tag.

	A	C	G	I	M	W
'no person'	-	2	-	-	-	-
'person'	597	-	85	589	597	596
'human'	597	-	20	-	-	-
'people'	596	587	-	45	-	376
'people (face)'	-	-	-	-	-	134
No humanness tags (H')	0	12*	492	8	0	1

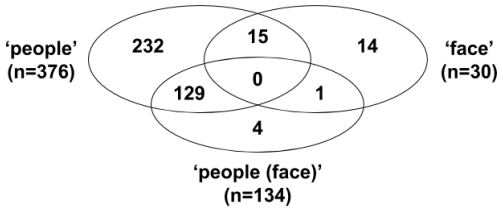

Figure 3: Co-occurrence analysis of Watson's three tags.

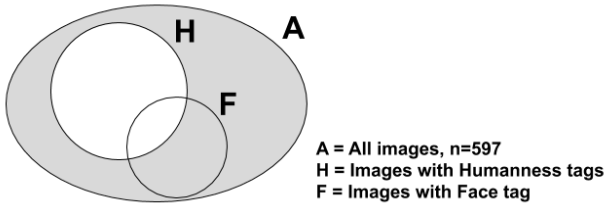

Figure 4: Venn Diagram demonstrating H' (the shaded area), the set of images which did not receive any humanness tags.

4.2 RQ2: Is everyone equally 'human'?

Only one ITA (Clarifai) explicitly declared there was 'no person' in the images, and it did so twice (Table 2); all other tags we identified were affirmative indicators of 'humanness.' Hereon, when we discuss 'images which did not receive any humanness tags' (i.e. H'), for Clarifai this will include the two images which *did receive* the 'no person' tag (unless indicated otherwise).

In a few cases the humanness tag in question was used on all images (Amazon's 'person' and 'human' tags, Microsoft's 'person' tag). However, most tags still had some margin of error, ranging from one image in 597 (Amazon's 'people' tag) to 577 images in 597 (Google's 'human' tag).

In addition to ITAs' errors regarding specific humanness tags, we also questioned whether there were images that did not receive any humanness tags at all (H'). In fact, with the exception of Amazon and Microsoft, all ITAs had at least one image where they failed to indicate the presence of a human. Watson only missed one image, while Clarifai missed six (eight when we include those tagged with 'no person'), and Imagga missed eight. Google, however, did not use any humanness tags for 492 images.

Table 3 shows the race and gender breakdown of the images that did not have any humanness tags. Relevant for the next analysis, each cell in this table has two numbers: the second number is the total number of images (of a particular social group) which did not receive any humanness tags (H'), while the first number indicates how many of those images did end up receiving a 'face' tag ($H' \cap F$). For example, there were five images of White women that Clarifai did not tag with any humanness tags, and two of those five images still received a 'face' tag.

In every ITA's case, and for every racial group, more images of women lacked humanness tags as compared to images of men. Overall, we also observed that (even within most ITA's outputs)

Table 3: Each cell: *n* of ($H' \cap F$) / H' for a particular ITA and social group. *Includes two images which received both the 'no person' and the 'face' tags.

	n	C	G	I	W
Asian Women	57	1/1	56/56	-	-
Asian Men	52	-	40/40	-	-
Black Women	104	1/2*	82/91	5/5	0/1
Black Men	93	1/1*	67/68	-	-
Latina Women	56	0/1	56/56	2/2	-
Latino Men	52	0/1	31/31	-	-
White Women	90	2/5	86/86	1/1	-
White Men	93	1/1	64/64	-	-
Total	597	6/12*	482/492	8/8	0/1

Table 4: Frequency of the 'face' tag, and the number of images lacking humanness and/or 'face' tags, for each ITA. *Includes the two images with the 'no person' tag.

	A	C	G	I	M	W
'face', F	597	466	586	597	57	30
No humanness tags, H'	0	12*	492	8	0	1
Lacking humanness & 'face' tags, $H' \cap F'$	0	6	10	0	0	1

higher proportions of Black people - particularly Black women - were receiving no humanness tags.

4.3 RQ3: 'Faces' without 'humanness'

All of the six ITAs examined had a 'face' tag; the frequency of use (i.e. *n* of F) is detailed in Table 4, along with the number of images that did not receive any humanness tags (H'), and those that lacked both the humanness and 'face' tags ($H' \cap F'$). The first row demonstrates that four out of the six ITAs audited correctly identified the face in most images, even though they are not services that are specialized in facial recognition.

Similarly, Table 3 shows that even when there are no humanness tags on an image, there may still be a 'face' tag ($H' \cap F$). In fact, that is the case for most images missing humanness tags. It is important to note here that the two images Clarifai tagged with 'no person' (one Black woman & one Black man) were also both tagged with a 'face' tag and are included in the counts in Table 3.

Lastly, the final row on Table 4 shows how many images lack both humanness *and* 'face' tags ($H' \cap F'$). Out of these 17 images, eleven (G: 9, C:1, W:1) depicted Black women, three (C) depicted White women, one (G) depicted a Black man, one (C) depicted a Latina woman and one (C) depicted a Latino man.

5 DISCUSSION

While dehumanization can appear differently in human-human interactions, image tagging algorithms may dehumanize by either failing to tag images of individuals with humanness tags in their vocabulary, explicitly using a tag claiming there is 'no person' in the image, or using tags to enable or reinforce dehumanizing treatment of people. We found that the six ITAs audited each had a unique vocabulary of tags to indicate the presence of a human, and that at least one image failed to receive most of these tags (with the exception of three tags from two services, applied to all 597 images). When a tag to identify a human exists in the vocabulary of an ITA, and the image of a person receives a False Negative error for that tag, it follows that the image of this person does not have the 'quality' (features, or characteristics) that the algorithm uses to decide if there is a 'human' in the image. This matches the definition of *attribute-based* [28] *mechanistic dehumanization*, which is the denial of core, universal, inborn qualities ('human nature') to some people [21]. The tagging of images depicting people with the 'no person' tag, as Clarifai did, is *explicit* [24] *metaphor-based mechanistic dehumanization*. This is similar to the use of the 'animal' or nonhuman primate tags[36,37] on images depicting *only* humans, which is *explicit metaphor-based animalistic dehumanization*.

The co-occurrence analyses of the 'humanness' tags show that certain tags that we may treat as interchangeable (such as 'person' and 'human') are in fact not interchangeable in the ITAs' model as they are applied to different images (e.g. Google). In some cases, the tags that were textually overlapping (which we would expect to see reflected in the outputs) were not always applied to the same images, indicating that the concepts were not explicitly related in the model's calculations (e.g. Watson). Therefore, if we are to shift from "maximizing prediction and accuracy" to truly understanding the context, impact, and potential harm of an algorithm [8], we must ask what the calculations are *really* taking into consideration, and whether we trust that these calculations align well enough with the real, social definitions of these concepts to use in applications which may affect people's lives.

In fact, each of these types of errors – regardless of how many other successful predictions the ITA has made – will affect one person's life when used to make decisions about that person in an application. How *other* people *tend* to be treated by a system (i.e. 'accuracy' or 'significance'; whether belonging to the same social group, or overall) does not change an error's negative effects on a person. For example, if the image with the error is in an algorithm which ultimately results in some benefit, the dehumanized person may miss the chance to receive this benefit; this is often the case in biased models dealing with non-visual data as well.[38,39] In applications such as autonomous vehicle parking,[40] the failure to detect a human [51] can create the risk of physical harm. Even in seemingly "harmless" cases where an image is automatically tagged to enable quick categorization on an image-based social network, tags have the potential to create harm.[41,42] In none of these cases will the 'target' of dehumanization receive less harm, or even feel better,

[36] https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos
[37] www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas
[38] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
[39] https://www.theverge.com/2015/7/7/8905037/google-ad-discrimination-adfisher
[40] https://www.clarifai.com/blog/clarifai-featured-hack-val.ai-is-a-parking-app-for-your-self-driving-car
[41] https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos
[42] www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas

just because the algorithm made many more successful predictions relative to its errors. In the principles suggested by Birhane and Cummins, to move away from algorithmic injustice, we need to start by "cent[ering] those disproportionately impacted and not solutions that benefit the majority" [8]. Regardless of what the tests on an algorithm may say about their 'overall accuracy' on hypothetical datasets, we must keep questioning whether they are accurate in the way we need them to be in a certain context,[43] and how the (however small) error rates may affect real human lives.

Of course, just as harmful social biases in human-human interactions are directed towards out-groups [20], technology's harmful biases are usually found to concentrate on marginalized communities (who are often the out-groups to those creating and enforcing technological 'solutions'[44]). Our results indicate this kind of bias may also exist in ITAs, as we observed more errors for women and Black people, peaking for Black women who are at the intersection of these two marginalized identities. While examining the True Positive rate for their own analyses, Scheuerman et al. theorize that a lower True Positive rate on a tag indicates that the ITAs' 'understanding' of that concept in its model is "more specific and bounded" than other tags with higher True Positive rates [45]. Our results may then indicate that the calculations for 'humanness' are delimited in ways that exclude some people who are women and/or Black, which implies the 'understanding' of humanness in the ITAs' models may be centering White men. Our results also suggest that in cases where some people are 'dehumanized' by the ITA and blocked from receiving benefits, they are likely to be from already-marginalized backgrounds; so in allowing these errors to persist, the providers of ITAs may be deepening the structural inequality in society.

As discussed in Section 2.3, dehumanization is not only the explicit person/no person declaration that an ITA may make about an individual in an image; it is also the very logic behind the systems that claim to be able to represent humans in numbers and vectors and make predictions regarding their abstract qualities. The identification of faces is the foundation for these predictions of abstract qualities as well as surveillance, often prioritizing profit or efficiency over the person's privacy or well-being.[45] In this light, our findings for RQ3 – that some people's faces are still be recognized by some ITAs even when their 'humanness' isn't – appear to demonstrate that people are often seen as data sources before they are seen as complete human beings (which exemplifies mechanistic dehumanization/objectification). Finally, there were some images lacking both humanness and face tags, indicating a complete erasure of the individuals depicted. Perhaps unsurprisingly, most were Black women and other people of color, illustrating the social groups most unlike the concept of 'human' as encoded in the ITAs' models.

[43] https://news.sky.com/story/met-polices-facial-recognition-tech-has-81-error-rate-independent-report-says-11755941

[44] https://www.revealnews.org/article/heres-the-clearest-picture-of-silicon-valleys-diversity-yet/

[45] https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

5.1 Limitations and future work

We used a dataset that was created with the live versions of the ITAs in October 2018 [3]. At the beginning of 2020, Google removed gendered tags (e.g. 'man,' 'woman') from its corpus,[46] replacing them with the 'person' tag. Therefore, **our results for Google do not apply to their currently-live ITA**. The other ITAs may have also been updated since the dataset creation, however there is generally very little transparency regarding what is updated and when. For example, shortly before this paper was published, Clarifai released a model specifically for human detection,[47] which may be tested in a future work along with the general model. In addition, while the controlled dataset of images allowed us to systematically compare the treatment of social groups by ITAs, the images do not reflect the real use cases where images depict people in different manners.

Future work may consider other tags that indicate humanness indirectly (like 'woman,' 'boy,' 'person with mohawk'), analyze the confidence scores along with the tags of interest, or look into the use of tags relating to body parts (like nose, mouth, eyes). The Google ITA for example, which had large number of images with no humanness tags, had many other tags describing body parts at the time of dataset creation.

6 CONCLUSION

Computer vision applications 'looking at people' (i.e. when applied on images of people) have social and legal significance [17]. The underlying logic of this image processing is to take a group of colored data points – which has already reduced three dimensions of the physical world into two – and claim that granular calculations on such data can reveal 'hidden truths' or non-visual facts about the individuals in the image. This information is then uncritically used to feed as inputs into other algorithms, the outputs of which affect people's lives.

We've found that image tagging algorithms do not recognize the 'humanity' in everyone, much less for people belonging to marginalized groups. The omission of certain tags relating to 'humanness' can lead to dehumanization, with large-scale negative effects on many people, impacting the well-being of society. Furthermore, even when such disparity in errors (or the overall accuracy of the ITA) is improved, the outputs can still be used in dehumanizing ways. Therefore, it is important for every system to consider the social context that it will be implemented in, drawing on experiences and research about existing structural inequality.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and under Grant Agreement No 810105 and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

[46] https://www.businessinsider.nl/google-cloud-vision-api-wont-tag-images-by-gender-2020-2/

[47] https://www.clarifai.com/models/people-detector

REFERENCES

[1] Jens Allwood. 2017. Is Digitalization Dehumanization?—Dystopic Traits of Digitalization. *Proceedings* 1, 3 (2017). https://doi.org/10.3390/IS4SI-2017-04120

[2] Peter Asaro. 2012. On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross* 94, 886 (2012), 687–709.

[3] Pınar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2019. Social B(eye)as: Human and Machine Descriptions of People Images. *Proceedings of the International AAAI Conference on Web and Social Media* 13, 01 (Jul. 2019), 583–591. https://ojs.aaai.org/index.php/ICWSM/article/view/3255

[4] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671–732. Issue 3.

[5] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest* 20, 1 (2019), 1–68. https://doi.org/10.1177/1529100619832930

[6] Brock Bastian and Nick Haslam. 2011. Experiencing dehumanization: Cognitive and emotional effects of everyday dehumanization. *Basic and Applied Social Psychology* 33, 4 (2011), 295–303.

[7] John J Beckers and Henk G Schmidt. 2001. The structure of computer anxiety: A six-factor model. *Computers in Human Behavior* 17, 1 (2001), 35–49.

[8] Abeba Birhane and Fred Cummins. 2019. Algorithmic Injustices: Towards a Relational Ethics. arXiv:1912.07376 [cs.CY] Presented at the Black in AI workshop at NeurIPS 2019.

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[10] David Danks and Alex John London. 2017. Algorithmic Bias in Autonomous Systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, Australia) *(IJCAI'17)*. AAAI Press, 4691–4697.

[11] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE.

[12] Ronald J Deibert. 2013. *Black code: Inside the battle for cyberspace*. Signal.

[13] Lina Dencik, Arne Hintz, Joanna Redden, and Emiliano Treré. 2019. Exploring Data Justice: Conceptions, Applications and Directions. *Information, Communication & Society* 22, 7 (2019), 873–881. https://doi.org/10.1080/1369118X.2019.1606268

[14] Hamid Ekbia and Bonnie Nardi. 2014. Heteromation and its (dis)contents: The invisible division of labor between humans and machines. *First Monday* 19, 6 (May 2014). https://doi.org/10.5210/fm.v19i6.5331

[15] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[16] Francis Galton. 1879. Composite Portraits, Made by Combining Those of Many Different Persons Into a Single Resultant Figure. *The Journal of the Anthropological Institute of Great Britain and Ireland* 8 (1879), 132–144.

[17] Jake Goldenfein. 2019. The Profiling Potential of Computer Vision and the Challenge of Computational Empiricism. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. ACM, New York, NY, USA, 110–119. https://doi.org/10.1145/3287560.3287568

[18] Jodi Halpern and Harvey M Weinstein. 2004. Rehumanizing the other: Empathy and reconciliation. *Hum. Rts. Q.* 26 (2004), 561.

[19] Omar Sultan Haque and Adam Waytz. 2012. Dehumanization in medicine: Causes, solutions, and functions. *Perspectives on psychological science* 7, 2 (2012), 176–186.

[20] Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and social psychology review* 10, 3 (2006), 252–264.

[21] Nick Haslam, Stephen Loughnan, Yoshihisa Kashima, and Paul Bain. 2008. Attributing and Denying Humanness to Others. *European review of social psychology* 19, 1 (2008), 55–85.

[22] Nick Haslam, Stephen Loughnan, Catherine Reynolds, and Samuel Wilson. 2007. Dehumanization: A New Perspective. *Social and Personality Psychology Compass* 1, 1 (2007), 409–422.

[23] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 88 (Nov. 2018), 22 pages. https://doi.org/10.1145/3274357

[24] Nour Kteily, Emile Bruneau, Adam Waytz, and Sarah Cotterill. 2015. The ascent of man: Theoretical and empirical evidence for blatant dehumanization. *Journal of personality and social psychology* 109, 5 (2015), 901.

[25] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. *Proceedings of the International AAAI Conference on Web and Social Media* 13, 01 (Jul. 2019), 313–322. https://ojs.aaai.org/index.php/ICWSM/article/view/3232

[26] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[27] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[28] Steve Loughnan, Nick Haslam, and Yoshihisa Kashima. 2009. Understanding the Relationship between Attribute-Based and Metaphor-Based Dehumanization. *Group Processes & Intergroup Relations* 12, 6 (2009), 747–762. https://doi.org/10.1177/1368430209347726

[29] Steve Loughnan and Maria Giuseppina Pacilli. 2014. Seeing (and treating) others as sexual objects: toward a more complete mapping of sexual objectification. *TPM: Testing, Psychometrics, Methodology in Applied Psychology* 21, 3 (2014).

[30] Helen Nissenbaum Lucas D. Introna. 2000. Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society* 16, 3 (2000), 169–185. https://doi.org/10.1080/01972240050133634

[31] Caitlin Lustig and Bonnie Nardi. 2015. Algorithmic Authority: The Case of Bitcoin. In *Proceedings of the 2015 48th Hawaii International Conference on System Sciences (HICSS '15)*. IEEE Computer Society, Washington, DC, USA, 743–752. https://doi.org/10.1109/HICSS.2015.95

[32] 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 (01 Dec 2015), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

[33] Martha C Nussbaum. 1999. *Sex and social justice*. Oxford University Press.

[34] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, NY.

[35] 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. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 106–114. https://ojs.aaai.org/index.php/HCOMP/article/view/5267

[36] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA) *(KDD '08)*. ACM, New York, NY, USA, 560–568. https://doi.org/10.1145/1401890.1401959

[37] Seeta Peña Gangadharan and Jędrzej Niklas. 2019. Decentering technology in discourse on discrimination. *Information, Communication & Society* 22, 7 (2019), 882–899.

[38] Tawana Petty, Mariella Saba, Tamika Lewis, Seeta Peña Gangadharan, and Virginia Eubanks. 2018. Reclaiming our data: interim report, Detroit. https://www.odbproject.org/wp-content/uploads/2016/12/ODB.InterimReport.FINAL_.7.16.2018.pdf. Last accessed 27-Apr-2021. Our Data Bodies.

[39] Kathleen H. Pine and Max Liboiron. 2015. The Politics of Measurement and Action. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3147–3156. https://doi.org/10.1145/2702123.2702298

[40] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 429–435. https://doi.org/10.1145/3306618.3314244

[41] Lorna Roth. 2009. Looking at Shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity. *Canadian Journal of Communication* 34 (2009), 111–136.

[42] Laurie A Rudman and Kris Mescher. 2012. Of animals and objects: Men's implicit dehumanization of women and likelihood of sexual aggression. *Personality and Social Psychology Bulletin* 38, 6 (2012), 734–746.

[43] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability. *International Journal of Social Robotics* 5, 3 (01 Aug 2013), 313–323. https://doi.org/10.1007/s12369-013-0196-9

[44] 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.

[45] 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.

[46] Juliana Schroeder and Nicholas Epley. 2016. Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General* 145, 11 (2016), 1427.

[47] Clay Shirky. 2009. A Speculative Post on the Idea of Algorithmic Authority. https://web.archive.org/web/20191030063204/http://www.shirky.com/weblog/2009/11/a-speculative-post-on-the-idea-of-algorithmic-authority/

[48] C.W.J. van Miltenburg, Desmond Elliott, and P.T.J.M. Vossen. 2018. Talking about other people: an endless range of possibilities. In *Proceedings of the 11th International Conference on Natural Language Generation*. International Natural Language Generation Conference (INLG), 415–420.

[49] Emiel van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset. arXiv:1605.06083 [cs.CL]

[50] Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology* 114, 2 (2018), 246.

[51] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive inequity in object detection. arXiv:1902.11097 [cs.CV]

[52] Xiaolin Wu and Xi Zhang. 2016. Automated Inference on Criminality using Face Images. *ArXiv* abs/1611.04135 (2016).

[53] Ting Zhang, Ri-Zhen Qin, Qiu-Lei Dong, Wei Gao, Hua-Rong Xu, and Zhan-Yi Hu. 2017. Physiognomy: Personality traits prediction by learning. *International Journal of Automation and Computing* 14, 4 (2017), 386–395.