Marine CHEFDEVILLE
Karolína HYBSKÁ
Luísa FRANCO MACHADO
Victoria SHIVENKOVA

Decoding Biases in AI - Spring Semester 2021
Final Group Assignment

Investigating Racial Bias in Artificial Intelligence Systems

The Case of Google Cloud Vision API

Table of Contents

1. Introduction

As artificial intelligence and machine learning technologies continue to transform the manner in which the world functions, fascination is being gradually replaced by skepticism as consequences remain unknown. While the development of these emergent technologies is progressing with rapid speed, their principal issue remains: potential reproduction of innate bias. Currently, image recognition and classification is a hot topic in the field of AI as developers strive to build better and more advanced algorithms capable of understanding complex visual content.

1.1. What is Image Detection?

Image detection is a technology whereby a computer processes an image and detects the objects in it. Additionally, this technology sometimes offers the possibility to classify the items detected in the image (image classification), typically by labelling the objects within, allowing them to be sorted into "classes".

The intent behind image detection AI and ML applications is to imitate human interactions, in this case, *vision*. If you consider, the human brain carries out these processes of object detection and recognition every second of every day for split seconds. When a human views an image of a car, it detects the shape of the car instantaneously, alongside other adjacent information. AI and ML systems attempt to complete the same task, understanding images and videos (Paramvir Singh 2018). The most common uses of AI/ML systems is to detect **what** is in an image: **Face recognition**, **Detecting objects**, **Text detection**, and **Logos**, **landmark detection** (*idem*.).

Machine learning and AI is progressively becoming a focal player in image detection and classification. However, this development is not without its challenges as these systems require significant training to carry out this task. Developers must "feed" the AI system with labelled data - mounds of images containing various pieces of information, the objects, the location, the labels. These images must be varied enough to show the object at different distances, in different locations, in different lighting conditions, etc, as the machine learning algorithms must be capable of correctly classing the object even with these changes in environment.

These advancements are increasingly significant as users in the digital world are perpetually equipped with high-end cameras in their pockets, allowing them to produce and consume high quantities of visual content (Paramvir Singh 2018).

1.2. What is Google Vision API?

Google Vision API is one of these attempts to undertake the mammoth task of image detection. According to its website, Google Vision is "*a pre-trained model that detects objects and faces, landmarks, performs image recognition, optical character recognition*

(OCR), classification, labelling, and text extraction of printed or handwritten text from images" ("Cloud Vision" 2019). Additionally, according to Paramvir Singh (2018), ***Cloud Vision API*** *enables developers to understand the content of an image by encapsulating powerful machine learning models in an easy-to-use REST API. It quickly classifies images into thousands of categories (such as, "sailboat"), detects individual objects and faces within images, and reads printed words contained within images. You can build metadata on your image catalogue, moderate offensive content, or enable new marketing scenarios through image sentiment analysis."*. Once more, there is an emphasis on the capabilities of this software to assign a vast array of labels to an image, to detect faces, animals, brand logos, and landmarks (with remarkably accuracy), to detect text, ascertain the possibility of explicit or violent content and lastly, to execute a Google search to establish a relationship between the given image and images available online.

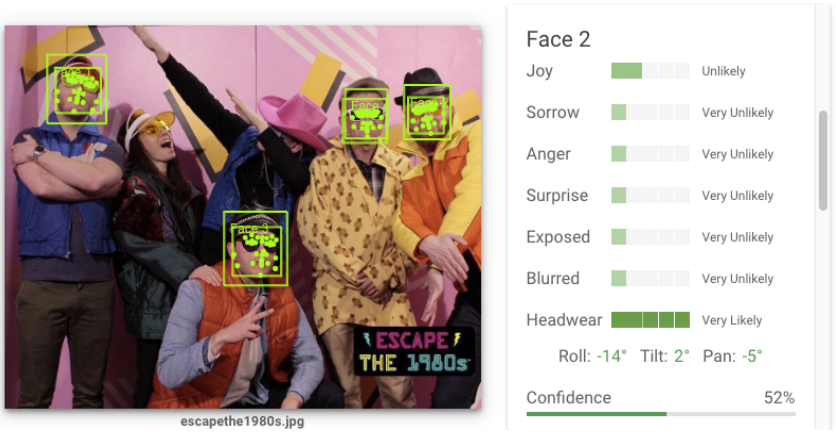
Having said this, Google Vision API is not alone in its endeavour to label the world's imagers. Amazon's *Rekognition*, IBM's *Watson Visual Recognition* and *Clarifai* are but a few of the diverse visual detection systems out there (Altex Soft 2019). Rekognition offers a cloud-based platform which provides two types of analysis: pre-trained algorithms and customizable algorithms that a user can create using their own dataset (Amazon 2021). IBM Watson has been developed with deep learning algorithms for image analysis (IBM 2020). It allows users to build, train and test their own models with the Watson Studio (IBM 2020). Clarifai operates through 14 pre-built computer vision models which return predictions based on the imputed image (Clarifai 2021).

Ultimately, we settled on analysing Google Vision API because we launched our project through a curiosity of how Google Photos operates its face and landmark recognitions and in what manner it attaches labels to photos to allow for the search function. The research into Google Photos and its visual detection system led us to Google Vision API which seemed to be an ideal platform to analyse due to the user-friendly interface and the explicit labelling mechanism it comes with.

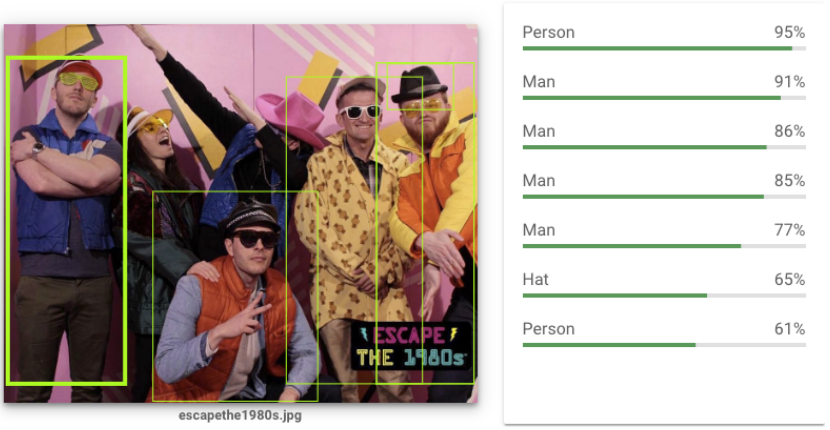
1.3. How do the features of Google Vision API work?

Now, we will briefly present how the features of the API work:

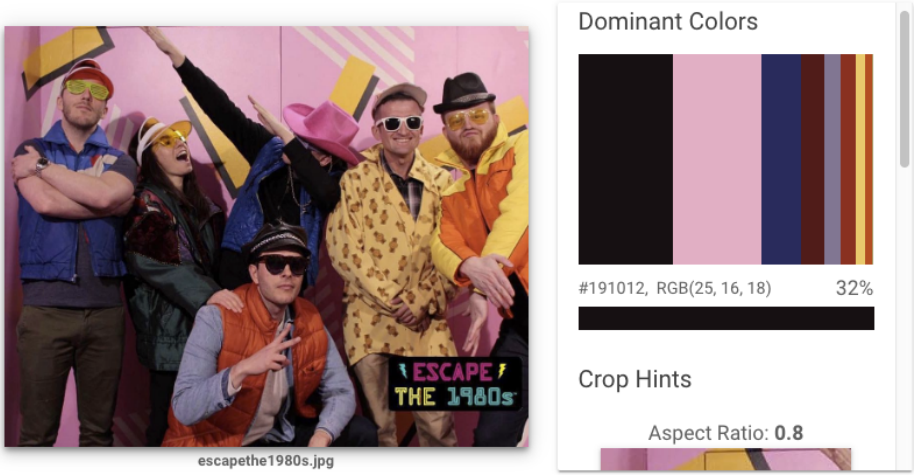
(1) **Face Detection.** The API performs a facial detection on local image files ("drag and drop") and remote images through URLs. When coding the API's instructions through Python, DETECT_FACES and DETECT_FACES_URI codes can perform multiple-face detection within an image, along with the associated key **facial attributes**, such as **emotional state** and **headwear** (Cohen 2020).



Similarly, the API can detect multiple objects within an image using the module Object localisation (identifies information about the object, its positionings and highlights it with a rectangle).

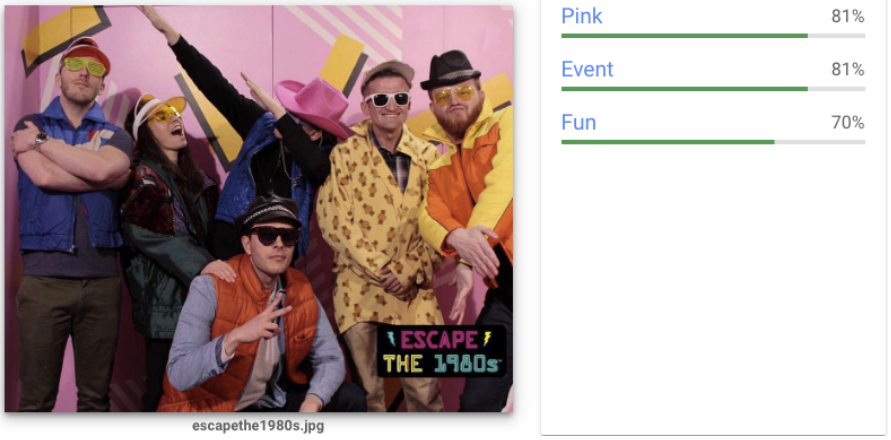


(2) **Image Attributes.** The API is able to detect the general attributes of the image, such as dominant colours and appropriate crop hints.



(3) **Label Detection.** The API can detect and extract information about objects in an image and attach a vast array of labels to them across a broad group of categories. Labels can identify human beings, animals, general objects, landmarks and monuments, activities, brand products and logos, even celebrity faces.

(4) **Text Detection (OCR).** The API's text detection can extract text from images

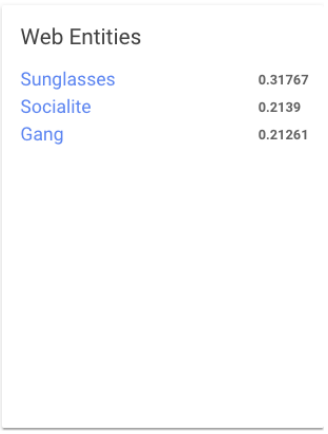


using the functions TEXT_DETECTION and DOCUMENT_TEXT_DETECTION annotations which install and support OCR (Cohen 2020).

escapethe1980s.jpg

(5) **Web Detection**. Web Detection detects web references to an image and locates it in cyberspace. It allows the user to use the given labels as criterias for the search and returns images and/or pages with full and partial matching images.

escapethe1980s.jpg



(6) **Explicit Content Detection**. The API is decently accurate when detecting explicit or violent content, such as adult content or violent content, within an image. This feature uses five categories (adult, spoof, medical, violence, and racy), and returns the likelihood that each is present in a given image. However, its notions of explicit content are varied.

2. Literature Review

2.1. Mislabelling of Images Using Google Vision API

The Google Vision API has been tried, tested and redeveloped multiple times in the past. Except the infamous label of "gorillas" being assigned to a group of people with darker skin, which Google solved by simply removing the term *gorilla* from its labelling vocabulary, there have been other rather inappropriate, either racists or gender-related, hiccups on the side of Google Vision API (Simonite, 2018).

In general, the rate of mislabelling is consistently higher for women as well as for people with darker skin. An example, very relevant for the current pandemic situation, is related to a thermometer, which is being used on a daily basis nowadays. The Google Vision Cloud assigned dramatically different labels for a light-skinned hand holding a thermometer than for a dark-skinned hand[1]. While the light-skinned hand was considered holding a monocular, the dark-skinned hand was labelled as holding a gun. The disparity in outcomes is caused by the labelling of the training data (images). Since the training set of images probably included more dark-skinned people in scenes depicting violence, the algorithm was more likely to choose a label from a similar lexical field; in the same way that the algorithm would perceive dogs as wolves in a snowy background or cows as dogs when captured on the beach resulting from the fact that computer vision does not recognise images and objects in the same way as humans do. Furthermore, people working on the Google Vision API are predominantly white and male. That can also be one of the reasons why such issues arise and why computer vision companies (not only Google) are only now starting to put in place processes to test and report failures related to racial discrimination and the like.

Even though some of the examples of object misrecognition might seem unimportant, they can have real-world consequences. For instance, the US uses weapon recognition tools at multiple public areas such as schools, supermarkets or concert halls. Some of the Europe's police forces use automated surveillance, while a Sri Lankan man has been wrongly accused of participating in bombings based on a face recognition. The systems used in those instances, while different, are very similar to Google Cloud Vision (Kayse-Bril, 2020).

2.2. Google Vision API Algorithm(s)

Use of the Algorithms

Currently, Google's Vision API has many uses, for individuals as well as for companies (e.g. companies managing media catalogs). Amongst them are: *detecting and moderating explicit, adult or offensive content* through sentiment analysis as well as image recognition without having to spend money on manual moderation where the Vision's

[1] See Appendix 1.

Safe Search can evaluate the likelihood that the given picture contains this type of contents; *improving the referencing of websites* marketers and people in charge of Search Engine Optimization can use the API to understand how Google labels images and optimising their choice and use of illustrations to better fit users' choice of keywords; *recognizing items* for shopping apps since from a simple picture, the Vision API can be used to propose similar items and thus power shopping apps.

In our project, we are only focusing on the labelling function of the Google Vision API. However, the system offers multiple other features which can be used to analyse images and thus, it has to rely on several powerful algorithms (e.g. CNN, OCR). The main one of these is the Convolutional Neural Network Algorithm (CNN).

Convolutional Neural Network (CNN)

CNN is an artificial neural network aimed at image recognition, which, in the case of Google, is trained on millions of images in a supervised form. That is to say that Google uses labelled images to train its algorithm (Menon, 2017). In the first step of image recognition, image classification is executed by extracting important information and characteristics from the images. For instance, a set of pixels representing a cat is removed from the background since a significant variation in RGB pixel values is identified (Achin, 2018). CNN also breaks every image into smaller clusters of pixels, which are called filters. These are then compared by the algorithm to other pixels with similar patterns. The algorithm itself is, however, very complex and computes its calculations in several hidden - convolutional - as well as non-convolutional layers. The first layer is also the simplest one where high-level patterns, such as rough edges and curves are detected. The following levels include detection of more intricate shapes and objects, such as feathers, fur, skin, eyes, etc. Only in the later layers, faces and animals are recognised[2]. As already mentioned, the Google Vision algorithm is trained on labelled data, but during the first layers of the CNN all the filter values are randomised. That is why the initial predictions of the algorithm make little sense. Gradually, each algorithmic prediction at each layer is compared with the already-existing labelled data thanks to the CNN's error, or loss, function. The loss function enables the CNN to update its filter values and then start the process again. In an ideal scenario, each iteration would be more accurate than the previous one (Google Cloud Tech, 2018).

Face detection

Google Cloud Vision offers a facial recognition tool relying on a Representational State Transfer (REST) Application Programming Interface (API), acting as a messenger. Basically, the API takes the user's request, tells the system what to do, before bringing the response back to the user. In the case of Cloud Vision API, the machine categorises

[2] See Appendix 2.

images (*query*), then searches and detects particular features, and finally displays them (*response*). The tool is complex enough to detect multiple faces within the same image, and their respective features and characteristics. However, regarding privacy, it is important to note that Google Cloud Vision does not include a face recognition functionality.

Labelling

Cloud Vision API has the ability to annotate the given picture with labels, by detecting and extracting pieces of information regarding many categories such as items, places, activities, moods, animals, features. As already mentioned, the algorithm is based on supervised learning, which means that annotated images are used for its training. Even if no official list of labels has been released by Google, people have been referring to some lists of more than 20,000 labels that have been used when training the algorithm. However, the labelling tool is sensitive to pictures' quality, and characteristics such as rotation, noise, clarity, and these will influence the accuracy of the labelling process (Apte, 2019). There also exists a tool, Cloud AutoML Vision, allowing everyone to train a ML model for other labels, when more specific image information is needed (e.g. cat breeds).

Optical Character recognition (OCR)

Additionally, Google's Vision API is able to recognize text through a tool called Optical Character recognition (OCR), allowing a better understanding of the given image. This tool can be developed through two different algorithmic processes:

1 - Pattern recognition : the program is trained with texts of various fonts, shapes and sizes, and is then able to distinguish characters (printed or handwritten) from its experience.

2 - Feature recognition : here, the program does not rely on confrontation to various texts and its experience in recognizing them, but on the analysis of every character's features (curves, number of lines, angles).

This tool is mainly used to convert handwritten text into printed text (by creating a black and white image, clearly distinguishing the characters from their background), to decipher text with a camera (reading envelopes to sort mail, license plates on the road, checks information when depositing at the bank, etc.), or, just as Google's Vision API, in order to detect, recognize and translate text in an image.

TensorFlow

TensorFlow is an open-source platform dedicated to machine learning ; it offers tools, libraries and shared resources accessible for everyone. It can be used for programming or

solving complex numerical computations for instance (Dahl and Ersson, 2018). Google relies on it as its machine learning framework, mainly for its image recognition tools (Google Photos, Google Translate) designed for consumers. As compared to other open-source platforms, TensorFlow is particularly fast and efficient, and provides APIs for a number of programming languages.

Where is the algorithm lacking?

Although Google praises its Vision API, which "quickly classifies images into thousands of categories, detects individual objects and faces within images, and finds and reads printed words contained within images." studies have shown the vulnerability of its algorithm when confronted to poor quality images. Thus, "using images from ImageNet dataset, it has been found that adding an average of 14.25% impulse noise is enough to deceive the API" (Hosseini, 2017), just as it is the case with rotated images. In these two situations, the algorithm has difficulty in detecting faces and text or coming up with accurate labels. In addition, CNN is still not very efficient when it comes to processing temporal data (in videos for example), since it does not make the link between the groups of pixels, processing them independently while they are linked in time.

Finally, the image recognition algorithm is based on human work, and is thus necessarily biased: the algorithm learns from associations in the data it is trained on. A study based on Google Cloud Vision, Microsoft Azure Computer Vision, and Amazon Rekognition led to similar results regarding gender biases (Schwemmer, 2020). Indeed, the study considers the algorithm to be biased as soon as "it returns labels at different rates for different demographic groups", and it is exactly what the finding showed. The analysis of images of women resulted in three times the amount of labels related to physical appearance as compared to men[3], and were recognized at substantially lower rates, thus confirming the reflection of engineers' own biases.

3. Method of Analysis

3.1. Data collection and selected analytical approach

With regard to the method of analysis, the initial stage involved the image collection and annotation process. To begin with, using the Google Search engine and image stock websites, we collected 33 portrait style images (stored in Google Drive). Alongside these original 33, we further collected 54 sport images, 30 occupational images and 30 leisure images. These images were diverse, particularly in the racial characteristics of the people within. The aim at this point was simply to observe how the API would treat some of these photos so that we could narrow our image analysis to one of these groups.

[3] See Appendix 3.

As an example, during our initial research phase we uploaded two photos of a person holding an air rifle. The first photo was a white woman holding the rifle (as seen below) and through alteration we darkened her skin tone for comparative purposes.

The API returned a range of labels related to, "wind instrument" (78% confidence), "violin family" (76% confidence), "music" (69% confidence), "musical instrument (64% confidence), "musician" (64% confidence), etc.[4] Additionally, while the safe search function did not note anything specific, it did attribute that violence is "unlikely" . Considering the blatant inaccuracy of image detection in this photo and inspired by the literature findings of how race impacts object detection, using some mediocre photoshopping skills we altered the photo to resemble a darker-skinned person.

When it came to the altered photo, the API initially returned similar results in object detection, perceiving a "wind instrument" again but at 59% confidence[5]. However, the label detection was arguably more precise with the top labels being "Air Gun" (84% confidence, "Trigger" (77% confidence), "Shotgun" (74% confidence), "Shooting" (74% confidence). Yet, some unusual labels were also making an appearance, such as "Fun" (75% confidence). Nevertheless, "Musician" (55%) and "Wind Instrument" (54%) still made an appearance, albeit with lower confidence. Lastly, the safe search only showed a small change, now both the gauges for "violence" and "racy" were at the second level, "unlikely".

Although the evidence is inconclusive given the small scale of research, this notified us to the possible discrepancies in label attribution depending on the skin tone of the people in the images.

Our choice to focus on the portrait photos boiled down to a desire to bring our analysis down to the most simple layer: what labels does Google Cloud Vision API attribute to photos and could it be considered that labels are attributed on a racial basis?

[4] See Appendix 4
[5] See Appendix 5

As such, the aim of the project was to carry out a small-scale image analysis (restrained by several obstacles which are discussed in the next section) to analyse how the AI algorithm labels people of different races. To make this more feasible, the choice of portrait photography ensured that the algorithm was focused on their facial features and expressions because the backgrounds of the photos were neutral and clothing was inconspicuous.

Ultimately, from our folder of 33 portrait photos, we analysed 26 photos. The breakdown of these photos is as seen in the following figure:

Incidence Rate of Race	
Asian	4
Black	8
Indian	1
Middle Eastern	3
White	8
Incidence Rate of Sex	
F	15
M	11
Total N° Labels	**50**

The analysis was carried out using the API's demonstrative version online (due to technical difficulties in getting access to the platform to code directly). This made the process rather cumbersome as each individual image had to be uploaded to the platform, analysed and its features/labels noted in an Excel Sheet.

Lastly, although gender did not figure centrally into our analysis, it was certainly something that we were mindful of throughout our analysis with the understanding that the sex of the person could play a role in the algorithms attribution of labels. Subsequently, a further analysis on the sex and gender perceptions of the algorithm would add much to the ongoing literature on the subject.

3.2. Limitations of our analysis

We encountered several obstacles as we advanced with our research and analysis. Starting by selecting the photos to be used, as much as we tried to avoid reflecting our own biases in the sample selection, it is impossible to guarantee that our own personal thoughts on race and ethnicity were not reflected in the pictures chosen. Despite this clear limitation, we decided to continue with the classification as we understood that the value added by considering the gender aspect in our analysis was greater than the possible misinterpretations and biases we could potentially be adding to the search.

After the pictures were selected and manually classified by the researchers based on the perceived gender and race, we reached the most complex part of our research: the photo analysis. In this part, we encountered several obstacles for actually being able to use the API and add it to our Python/Google Collab notebook. If you are a new user, Google requires you to sign up for a free trial in order to use the tool. For this, you need to insert your credit card details as well as send a picture of your credit card and document to be analyzed by the company as part of an identity check before you can have access to the service. However, we were using a collaborative Google account created for the project in which we had mixed different contact information from the members of the group.

Therefore, when we encountered the identity check, it was impossible to prove our identity since we had no established identity but rather a mix of different identities. As a result, we could not access the coding file that would grant us access to the API in our notebook. Consequently, we had to find quickly a different alternative to being able to use the tool without needing to connect to the service.

As an alternative for the code, we decided to use the free version of the tool that is available online. With it, we were able to get the same results we would have gotten if we had followed the original procedure, but we had to develop our own data collection system. Although useful, it took us 5x the time it would have taken to collect the data if we had been able to use the code, thus limiting our remaining time for analysis as well as the sample size (which had to be lower than expected due to the limitations).

When it comes to the analysis of the pictures, there is a risk that our analysis of the results given by the API is biased for a number of reasons. Firstly, it could be biased as the results given by the API take into account not only the person itself but also the background, the lighting, their clothing, and that changes the types of labels we will get. Secondly, for our analysis we decided to bring up results that we considered to be the most relevant. However, what we consider relevant is biased by our own background and experiences, making it impossible to take unbiased paths of analysis. Ultimately, we understand that the limitations of our procedure are important to be recognized but they do not undermine our research. On the contrary, recognizing and sharing the obstacles is an important part of public interest studies as it allows for other researchers to collaboratively think of ways to overcome these difficulties and reach unique conclusions.

3.3. Ethical implications

As mentioned in the previous section, our own personal thoughts and feelings on the topic are unsurprisingly ingrained into our sample choice and analysis. This leads us to a bigger questioning on the ethical implications of our research and procedure, since we are precisely addressing (the lack of) AI ethics in this study.

Classifying the people in the pictures according to their “race” was an incredibly challenging task, as even the definition of race itself is problematic and has enormous nuances. As a result, we had to arbitrarily designate the people in our sample certain races based on what we, white western women, perceived them to be. Even if we tried to avoid following this procedure, we needed to merge different races together in order to find patterns among them and fill our analysis and this was the most obvious way to do it. Additionally, as we had initially thought of also taking gender into consideration to understand if this would also have an impact in the labels received, we also needed to classify the portraits based on the perceived gender of the person. However, we understand that gender expression and identity go beyond perceived gender stereotypes that could lead to inferring that one person identifies with a certain gender.

Beyond the ethics of arbitrarily separating people in race and gender groups, we also engaged in a deeper reflection that inquired: what is the meaning of decoding the API classifications? In other words, what were we looking for and why? Considering that we live in a profoundly racist society in which white people are complicit with racism (DiAngelo, 2018), what is the legitimacy of white people to look for racism in technology? These questions are some of the ethical dilemmas we encountered when addressing the topic that led us to keep a constant self-awareness of our position as we performed our analysis.

4. Features of Google Vision API and our Findings

4.1. Google Vision API labelling system

The API holds five features that allow for a deep analysis of the images: face detection, object detection, label detection, explicit content detection, logo/landmark detection, and text detection (Optical Character Recognition [OCR]). Face detection stands for seven fixed characteristics detected on each face present in the image: joy, sorrow, anger, surprise, exposed, blurred and headwear. The classification ranges from "very unlikely" to "very likely" to detect each of the characteristics in the faces detected. Object detection allows for a simple detection of the main objects in the image. Label detection is a feature that provides different labels that characterize the entire image (and not only the faces) and that have a confidence level higher than 50%. Explicit content detection, also presented as "safe search" in the API, presents different categories that could show that the image is unsafe for different reasons: "adult", "spoof", "medical', "violence" and "racy". This classification, too, ranges from "very unlikely" to "very likely". Subsequently, logo/landmark defines the possibility of the API to detect in the image either a logo or a famous monument or landmark. Finally, text detection allows the API to read over any text present on the image.

In this project, we decided to focus specifically on label detection. This choice was motivated by the intent to analyse how successfully the AI system detects diverse racial characteristics and what labels, if any, the system attaches to them. This could allow us to detect any bias in how the API perceives the different images in a more detailed way, since the number of labels provided is much larger than other categories.

4.2. Label Detection Findings

In this section, we will present our interpretation of the results found in our label detection analysis (Excel sheet in annex), which will address a range of factors including individual results, comparative results, and our interpretation of the system and its functionality.

Hair and headgear

The most notorious and systematic mislabeling we perceived in our results is the misattribution of labels that represented headgear (for example “wig” and “lace wig”) to afro hair. In one of the pictures, the label “artificial hair integrations” was assigned to a black woman with afro hair. In one of the portraits, a black woman with afro hair had the labels “microphone”, “audio equipment”, “music artist”, “entertainment” and “singer”. While we cannot conclude these came as a reflection of her hairstyle, we did notice that she had bright white lighting on her face, which might have led to a “stage-like” sentiment.

The API had very specific hair labels, including “crew cut”, “caesar cut”, and “buzz cut”. The “jheri curl” label, that represents this specific hairstyle, was curiously assigned to a white woman with curly hair. This was a controversial result, since this hairstyle is considered to be a symbol of resistance amongst black people, who reject the appropriation of the cut by white people. To the human eye, it is unclear whether the white woman has a jheri curl hairstyle or not as most of her hair is not present in the picture. Contrastingly, one black woman who effectively had a jheri curl hairstyle did not receive this label, perhaps because the hair blends in with the background.

Still regarding head elements, we noticed that women with headscarves had less labels than women without headscarves, while men with headscarves did not encounter the same issue. For one of the labels, “neck”, we noticed that it was mistakenly attributed to people with headscarves where no other parts of the body are visible except the face. We infer that the API assumed that *because there was a face, there should be a neck* and failed to take into account situations where the neck would not be visible in a portrait.

Beauty and beauty standards

Interestingly enough, the Google Vision API does not only assign descriptive labels to pictures, but subjective ones as well. For example, it asserts that some people are related to the themes of “beauty” or “modelling”, based on a portrait picture. While we can assume that these labels do not come from an analysis of the person’s conformity to beauty standards, the outcome remains interesting to study. In our sample, three women received the “beauty” label, without any clear racial bias. However, the “model” label was only given to a white, blue-eyed woman, while another picture with the “beauty” label depicted a black woman in a clear context of modelling (jewellery, elaborate gesture, make-up) and did not get the label. While it is complicated to come up with conclusions based on the small scale of our project, the overrepresentation of white women in the modelling industry surely led to a difficulty for the machine to recognize modelling when performed by women with non-white facial features.

Finally, since our sample was only made of portrait pictures, all of them received the “portrait photography” label, with similar confidence levels (around 60%) across the database. However, Google Vision API has another label, simply “portrait” that was not

systematically assigned to the pictures. The six pictures that did not mention it were only non-white people (3 black people, 1 Asian, 1 Indian and 1 from the Middle-East). Again, we can suspect a lack of representativity in the way the machine was trained.

Body parts

At first sight, collecting appearing facial features and body parts does not appear as a possibly controversial activity. But digging into this category of labels raises additional concerns on Google vision API inner biases. As a matter of fact, the "moustache" label appears not only on pictures of men wearing a moustache, but also on two pictures featuring black women smiling, possibly because of their pronounced nasolabial fold. Still regarding hair issues, two black women were assigned the "facial hair" label while none appeared, leading us to think the darkness of the skin confused the machine.

However, these observations should be nuanced by the lack of accuracy from Google Vision API when identifying facial features. Thus, only half of the people had the "nose" label, while other features such as "eyes", "throat", "head" or "hair" seemed to be randomly assigned, independently of racial differences.

Cultural labels

Surprisingly enough, a « tradition » label appeared on some of the portraits, applying to four people, only non-white (two black people, one Indian and one from Middle East). We wanted to include this specificity to our analysis, as the word "tradition" is loaded with a biased meaning. Indeed, we tend to refer to culture and tradition while speaking of practices and features coming from a non-white ethnicity, as if being white was the neutral reference. Indeed, from a white perspective, tradition refers to non-western religious and spiritual symbols (headscarf, bindi, specific jewellery), while it would not be used to describe a catholic priest for example. One of the pictures also features the label "temple", probably because the Middle Eastern wears a red headscarf. This once again proves the human bias lying behind Google Vision API algorithm, created mainly by western people.

Clothing

Clothing is one of the main components of someone's culture. Yet, it is analysed by Google Vision API under a purely western lens. As we have seen before, non-western clothes are labelled as "traditional", they are alienated, while western clothing codes are perfectly interpreted by the algorithm. A perfect illustration of this is the "formal wear" and "businessperson" labels, which only appear on people wearing western formal clothing (usually a suit, a blazer, a tie), and ignores other cultures' sense of "formal".

Other

Finally, the incidence of seemingly random attributions of miscellaneous labels grabbed our attention. The label "organ" was assigned to two black people and one asian person. It is unclear whether this refers to the musical instrument or to bodily organs, but none were present in the pictures analyzed. Next, the label "flesh" was assigned to seven people from different races (black, white, and middle eastern). The label "pleased" was assigned to a man, while "makeover" was assigned to one indian woman, a black woman with afro hair, and one white woman with afro-like hair. Finally, a black woman with afro hair was also labeled as "people in nature".

5. Final Conclusions and Remarks

In our analysis, we used the Google Cloud Vision API to understand how racial biases are present in machine learning and artificial intelligence algorithms. By analyzing a set of 26 portraits from people of different skin colors and ethnicities, we chose to concentrate our work on the precise study of label detection, allowing the user to obtain keywords characterizing the image. Indeed, our goal was to detect biases in the way the algorithm labels pictures featuring a variety of ethnicities.

Our main takeaways, although also biased by the small scale of our project, have to do mainly with black people features and western cultural bias. On the one hand, black hair was inaccurately labelled as headgear, or artificial hair, and black models would not be featured as such. On the other hand, the labels given by the algorithm presented a strong western bias, using the word tradition only for non-white people, and correlating their clothes to religion when white people were often described as wearing formal or business cothes. This can be explained by the lack of representation in the datasets used when training the algorithm. Finally, it is necessary to acknowledge the multiple biases that come together in these results, namely regarding gender; labels associated with a black woman portrait would also result in less accuracy than for a man of the same ethnicity.

We conclude that although the biases present in the algorithm are more subtle than originally expected, there are still some obscure parts of it that we cannot ignore. When it comes to the next steps of this research, we would aspire in the future to scale up our analysis and repeat the same process with a larger database, with images more complex than portraits, for example sports, events, occasions, etc. As our aim is to continue decoding biases in AI, we understand that there is a need to be constantly overseeing these mechanisms and publicly sharing our research with other scholars so the negative impacts of digital innovations can be identified and (hopefully) mitigated.

References

Achin, V. 2018. "Image Recognition using Google Vision API" [online].

Altex Soft. 2019. "Image Recognition APIs: Google, Amazon, IBM, Microsoft, and more" [online].

Amazon. 2021. "What is Amazon Rekognition? - Amazon Rekognition" [online].

Apte, A. and al. 2019. *Countering Inconsistent Labelling by Google's Vision API for Rotated Images*. Manipal Institute of Technology, India.

Clarifai. 2021. "Welcome". Clarifai Guide.

Cloud Vision. 2019. Vision AI | Derive Image Insights via ML | Cloud Vision API. Google Cloud.

Cohen, M. 2020. "Using the Vision API with Python". Google Codelabs.

Dahl, O. and Ersson, S. 2018. *Specialization of an Existing Image Recognition Service Using a Neural Network*. KTH.

DiAngelo, R. 2018. *White Fragility: Why It's So Hard for White People to Talk About Racism*. Beacon Press.

Google Cloud Tech. 2018. How Computer Vision Works [Last accessed: 09/04/2021].

Hosseini, H. et al. 2017. *Google's Cloud Vision API Is Not Robust To Noise*. Network Security Lab (NSL), Department of Electrical Engineering, University of Washington, Seattle, WA.

IBM. 2020. "Watson Visual Recognition - Overview.".

Kayser-Bril, N. 2020. Google apologizes after its Vision AI produced racist results [Last accessed: 01/04/2021].

Menon, N. G. 2017. All you need to know about Google Cloud Vision API [Last accessed 01/04/202].

Neves, A. and Lopes, D. 2016. "A Practical Study about the Google Vision API". 22nd Portuguese Conference on Pattern Recognition.

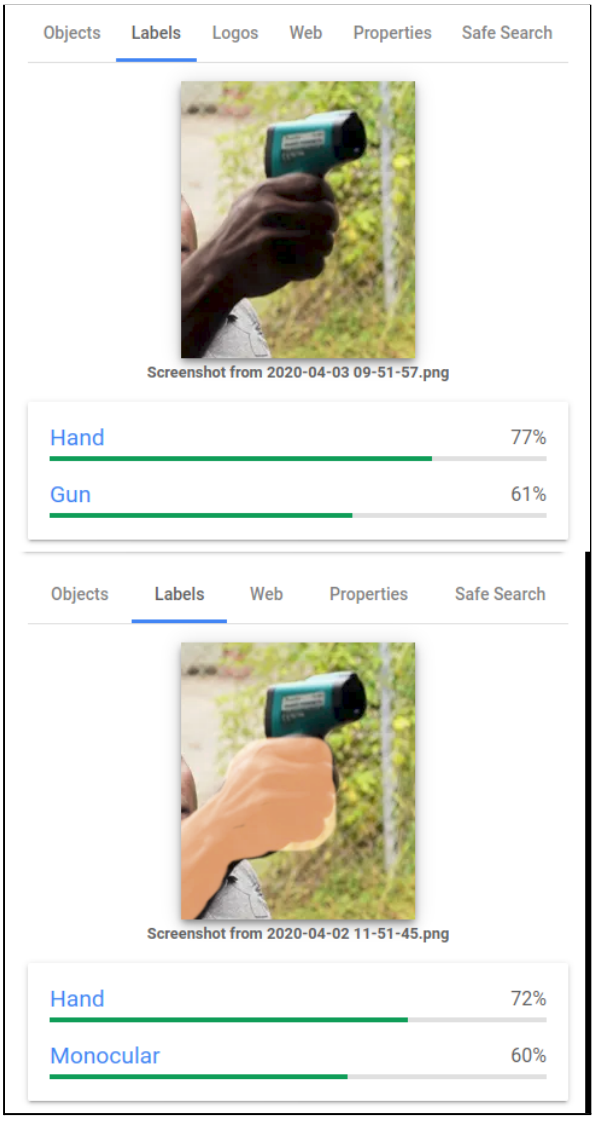
Paramvir Singh. 2018. "AI Capabilities in Image Recognition - towards Data Science." *Medium*.

Schwemmer, C. et al. 2020. *Diagnosing Gender Bias in Image Recognition Systems*. Socius.

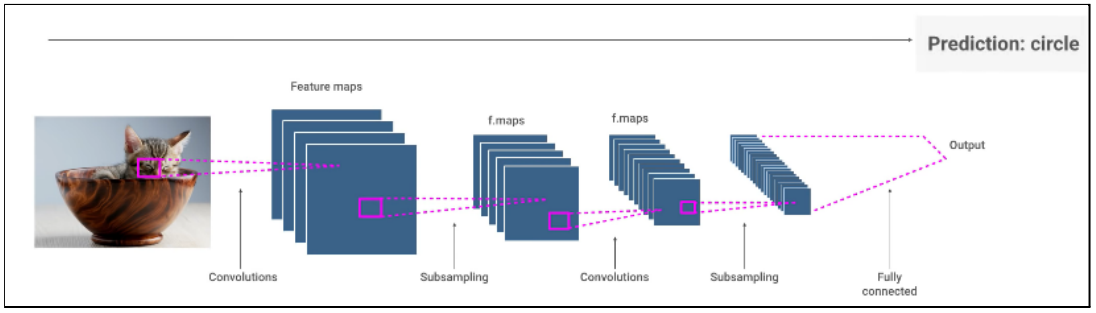
Simonite, T. 2018. When It Comes to Gorillas, Google Photos Remains Blind [Last accessed: 01/04/2021].

Appendices

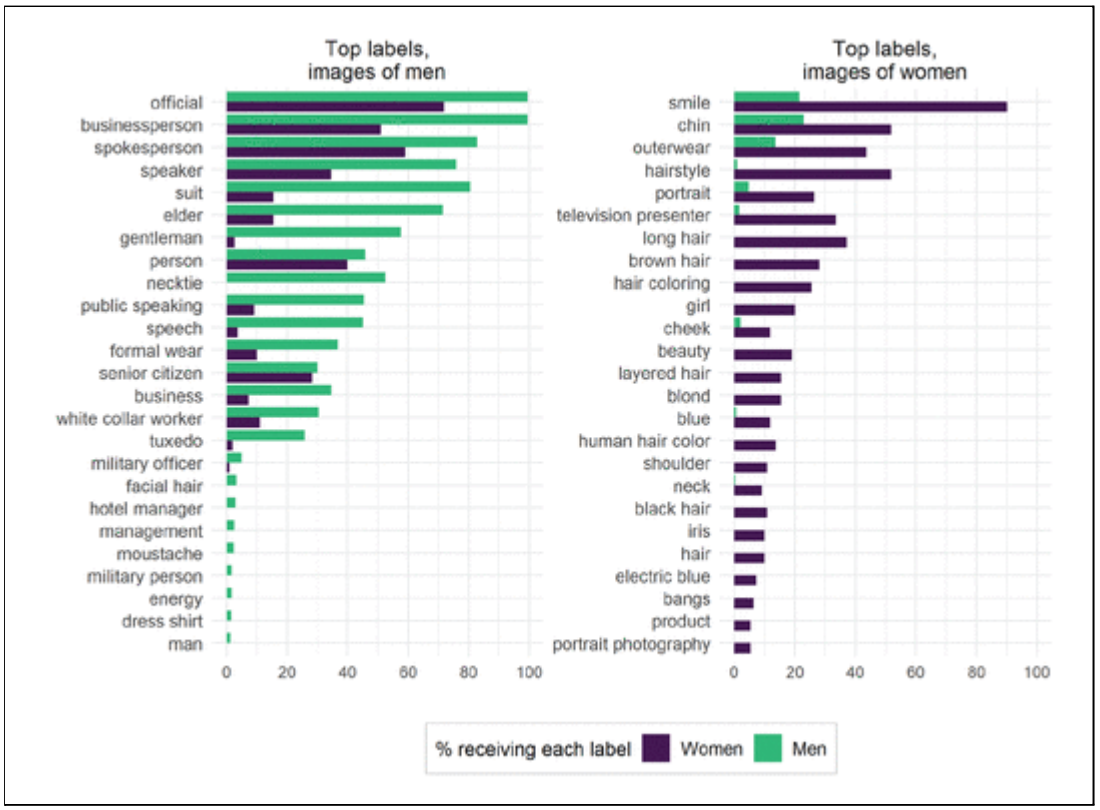
Appendix 1: Google Vision Cloud and the labels it assigned to the same images where the only difference was the skin colour (Kayser-Bril, 2020).



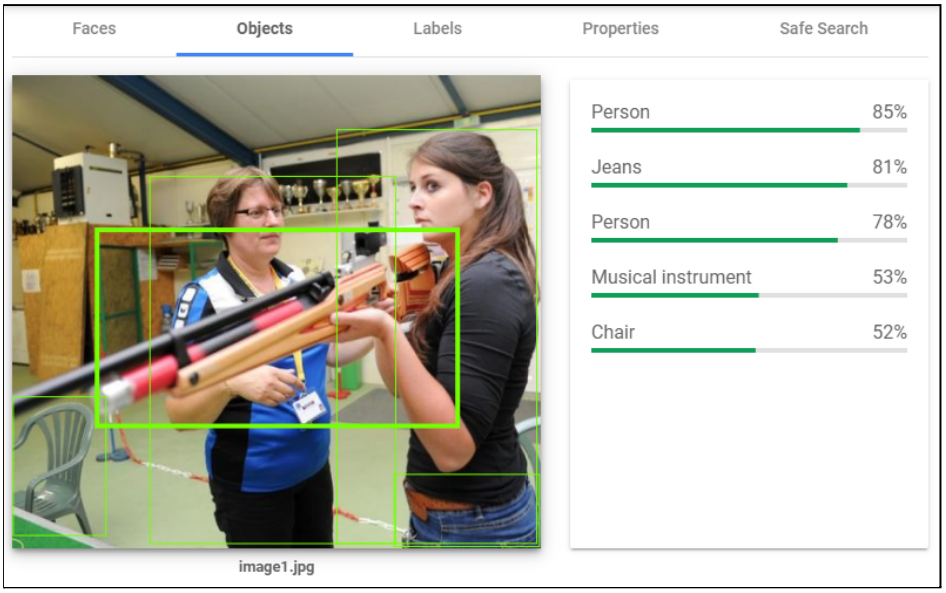
Appendix 2: Representation of the Convolutional Neural Network's layers during the image classification process (Google Cloud Tech, 2018).

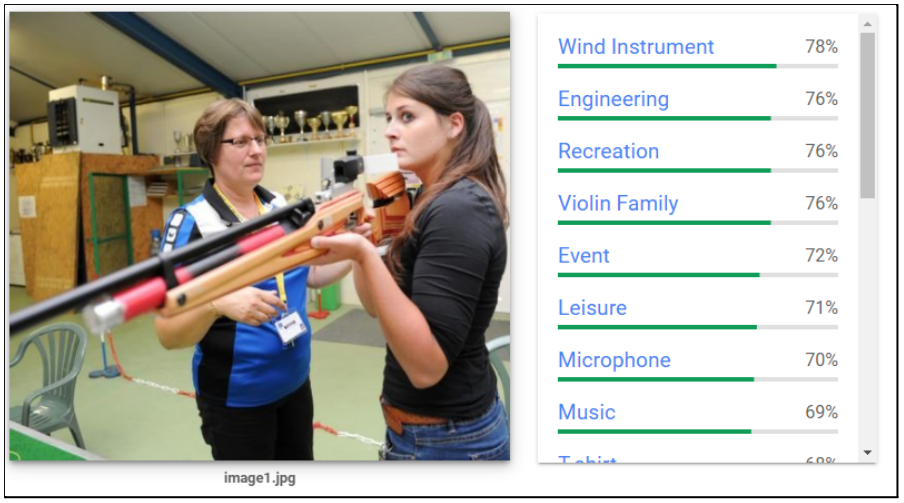


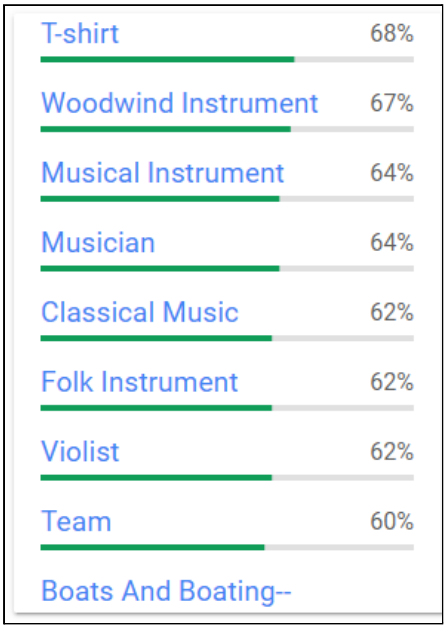
Appendix 3: Microsoft Azure Computer Vision labels applied to professional photographs of members of Congress (Schwemmer, 2020).

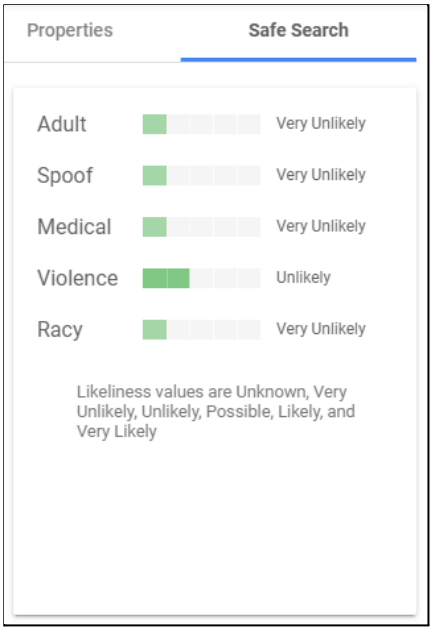


Appendix 4: Image Analysis of a white woman holding an olympic air rifle and the relevant labels (object detection; label detection; safe search)

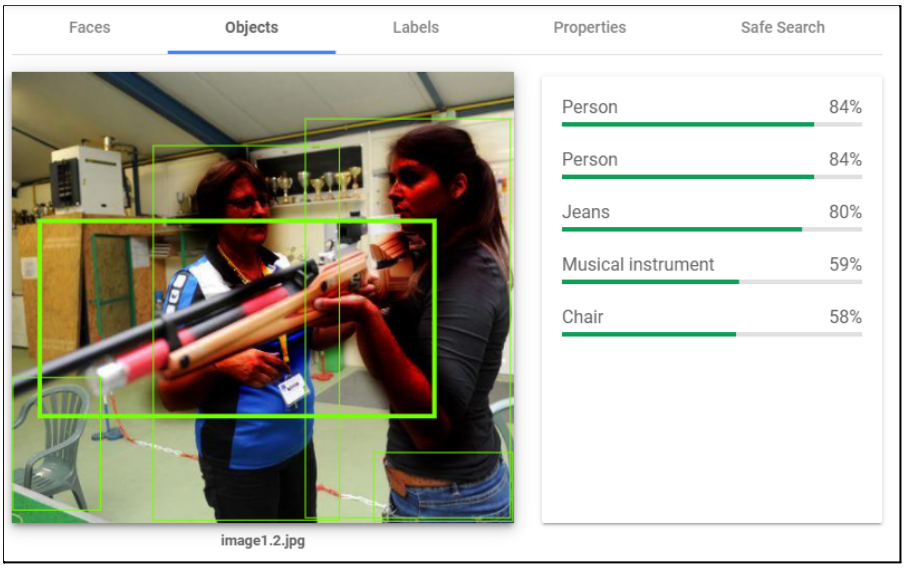








Appendix 5: Image Analysis of a photoshopped woman with darker skin tone holding an olympic air rifle and the relevant labels (object detection; label detection; safe search)



Faces Objects **Labels** Properties Safe Search

image1.2.jpg

Air Gun	84%
Trigger	77%
Fun	75%
T-shirt	75%
Shotgun	74%
Shooting	74%
Recreation	74%
Engineering	73%

Event	70%
Gun Barrel	69%
Eyewear	65%
Music	64%
Shooting Sport	63%
Personal Protective Equipment	61%
Leisure	60%

Games	57%
Musician	56%
Wind Instrument	55%
Fictional Character	54%
Belt	54%
Room	53%
Entertainment	52%
Team	52%

Faces Objects Labels Properties **Safe Search**

image1.2.jpg

Adult	Very Unlikely
Spoof	Very Unlikely
Medical	Very Unlikely
Violence	Unlikely
Racy	Unlikely

Likeliness values are Unknown, Very Unlikely, Unlikely, Possible, Likely, and Very Likely