National College of Art & Design.

Interaction design, School of Design.

Accountability and AI: the issues with algorithms used to censor explicit images online.

Diarmuid Farrell.

Submitted to the School of Visual Culture in Candidacy for the Degree of (BA) Interaction design, 2020

NCAD DUBLIN

National College of Art and Design

School of Visual Culture

I declare that this **Critical Cultures Research Project** is all my own work and that all sources have been fully acknowledged.

Signed:

Programme / department:

Date:

Table of Contents:

List of Illustrations:

Figure 7: Screenshot from Amazon Object and scene detection of Beauty Study_Che.

Available at:

(https://eu-west-1.console.aws.amazon.com/rekognition/home?region=eu-west-1#/label-detection)

Figure 8: Screenshot from Amazon Object and scene detection of Charles Bowan.

Available at:

(https://eu-west-1.console.aws.amazon.com/rekognition/home?region=eu-west-1#/label-detection)

Figure 9: Screenshot from Google Vision Api's analysis of Beauty Study_Che.

Available at: (https://cloud.google.com/vision/)

Figure 10: Screenshot from Google Vision Api's analysis of Charles Bowan.

Available at: (https://cloud.google.com/vision/)

Figure 11: Screenshot from Amazon Image moderation of Beauty Study_Che.

Available at:

(https://eu-west-1.console.aws.amazon.com/rekognition/home?region=eu-west-1#/image-moderation)

Figure 12: Screenshot from Amazon Image moderation of Charles Bowan.

Available at:

(https://eu-west-1.console.aws.amazon.com/rekognition/home?region=eu-west-1#/image-moderation)

Figure 13: Screenshot from Google Vision Api's analysis of adult content of Beauty Study_Che.

Available at: (https://cloud.google.com/vision/)

Figure 14: Screenshot from Google Vision Api's analysis of adult content of Charles Bowan.

Available at: (https://cloud.google.com/vision/)

Figure 15: Burstein, S. (2018) *Tumblr flagging footwear patent*

Available at: (*https://twitter.com/design_law?lang=en*)

Figure 16: Crusher, L. (2018) *Tumblr flagging women of colour*

Available at:

(https://lezlee-crusher.tumblr.com/post/180764866660/this-post-was-flagged-no-nudity-or-explicit)

Figure 17: Screenshot from Amazon Image moderation example image.

Available at:

(https://eu-west-1.console.aws.amazon.com/rekognition/home?region=eu-west-1#/image-moderation)

Figure 18: Yahoo (2018) *Open_nsfw example image*

Available at: (https://github.com/yahoo/open_nsfw)

Figure 19: Buolamwini, J. (2019) *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Introduction:

In the age of the algorithm, where more and more large institutions both corporate and governmental, utilize the power of AI to take broad sweeping actions that are affecting people's livelihoods become more and more prevalent, it's important to first understand and question how these systems operate and secondly, how accurate these systems are, because these systems are being used to make more important decisions for and about us. I will explore the issues with using such algorithmic actions through the lens of one of the most common uses of algorithmic tools online, the use of AI in censoring explicit imagery online. I will explore how these algorithms operate, how they are used and the issues that are built in to these systems at every stage and how these issues affect the decisions these systems make and the effectiveness of them.

Artificial intelligence systems and their use in image recognition are problematic. These systems have many biases and other systematic errors built into them, around how they are coded, the data they are trained on, how they classify objects in images as well as a plethora of other issues that are unknown due to the uncontrolled nature of how these systems operate. These issues that are at the core of how these systems work,and in turn create tools that are racist, sexist and exclusionary to groups that are overlooked and the most vulnerable in our society. (Ai Now institute, 2018; Goodman, 2017; Metz, 2019; Smith, 2019; Lohr, 2018, Buolamwini, 2017; IBM, 2020;)

Although I will be exploring this through the lens of its use within detection of explicit content in images and its subsequent use in censorship, a large body of work and research has already been put in to looking at the issues with facial recognition. I will explore this work into the issues in the systems built to recognize faces and analysing these findings in order to gain a more well rounded approach to some of the issues these systems have. I will also be exploring historical misuses of data and automated actions in an attempt to get a clearer idea of how these previous cases inform and affect our current algorithmic landscape.

In this paper, I will delve deep into one specific case study, where these large algorithmic actions by major corporations have directly affected the lives of their customers in a negative sense. This being Tumblr's use of an automated system to censor what they deem as “adult content” (Waterson, 2018) . I will explore the intentional and unintentional consequences of the use of this system and how it affected the people rely on the platform Tumblr provides. I will also try to explore the system Tumblr uses to censor these images to find issues that the system has and what could be the source of a number of errors that it caused.

To explore the topics that I will discuss in this paper I have utilized several methods of research. The primary focus of my research was using primary digital research (Rogers, 2013) , this involved writing image recognition scripts using common API scripts from leading companies in AI, to gain an understanding of how image recognition works as well as exploring machine learning in a general way. Along with this I explored well known and used cloud-based systems that have elements of image censorship, such as Google clouds safe search and Amazon AWS's Rekognition, to understand the language, operation and general workings of these systems. This primary research was then backed up by secondary research, in the form of papers and articles written on both the case study and theories to support my exploration into algorithmic accountability.

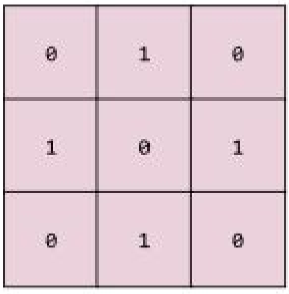
In the first chapter of this paper, I will explore the background of algorithmic systems and their use in image recognition, looking into how they operate and the theories behind how these systems take action. In the second chapter I will explore the case study, Tumblr's use of AI in censoring adult content, I will discuss how and why this was implemented as well as what the intentional and unintentional actions of this system were and how they affected their users. In the third chapter, I will explore some of the issues with both Tumblr's image recognition system as well as image recognition systems in general in terms of the language. This chapter will focus on the issues of the language these systems use in classifying the content of an image and classifying the nature of the explicit content. The fourth chapter will explore the issues with the data sources that these systems use in the training of these systems, and how the use of biased data affects the way in which these systems operate in terms of the predictions they make and their effectiveness.

Chapter 1: Background to AI and image recognition.

In order to build a frame of reference around some of the key terms, concepts, and systems I will be discussing in this paper, I will use this chapter to explain some of the key concepts and terminology within AI and Image recognition. I will use some general definitions to create a common understanding of the chapters ahead. In this chapter I will also discuss how common AI and image recognition systems work, highlighting examples from widely used pre-built systems as well as scripts I have experimented with as part of my research process.

First of all, It’s necessary to gain a common understanding of some of the language used and their definitions. When I am discussing the word AI or Artificial intelligence it refers to “A non-human program or model that can solve sophisticated tasks” (Google, 2020) for the purposes of this document I will be referring to the modern ideas of Artificial intelligence which often referred to as Machine learning, which refers to “A program or system that builds (trains) a predictive model from input data. The system uses the learned model to make useful predictions from new (never-before-seen) data drawn from the same distribution as the one used to train the model” (Google, 2020). Both the terms Machine learning and Artificial intelligence are often used interchangeably and for this document I will be referring to these terms interchangeably.

Now that the two main general terms in AI are defined I will go into more detail about how some of these artificial intelligence systems work, specifically those used in image recognition. The most common and well-known models in modern machine learning are called Neural networks, these are machine learning models which take inspiration from how the brain works, they are composed of a number of layers, each of which contains a series of connected nodes or neurons which are used for processing a set of inputs or outputs of the neurons. From an image recognition point of view, the specific kind of neural network used is called a convolutional neural network or CNN, these are “specifically designed to process pixel data” (Google, 2020). This model works by splitting up the processing of pixels into smaller chunks so it is easier to analyze. It typically operates on 3 layers, the convoluted layer which determines the pattern in which the matrices will be divided. See figure 1 for an example.



0	1	0
1	0	1
0	1	0

Figure 1. Convoluted Layer example

Then there's the pooling layer, otherwise known as spatial pooling, this divides the work up into smaller matrices that operate on input data. See figure 2 for an example.

Input data:

5	3	1
8	2	5
9	4	3

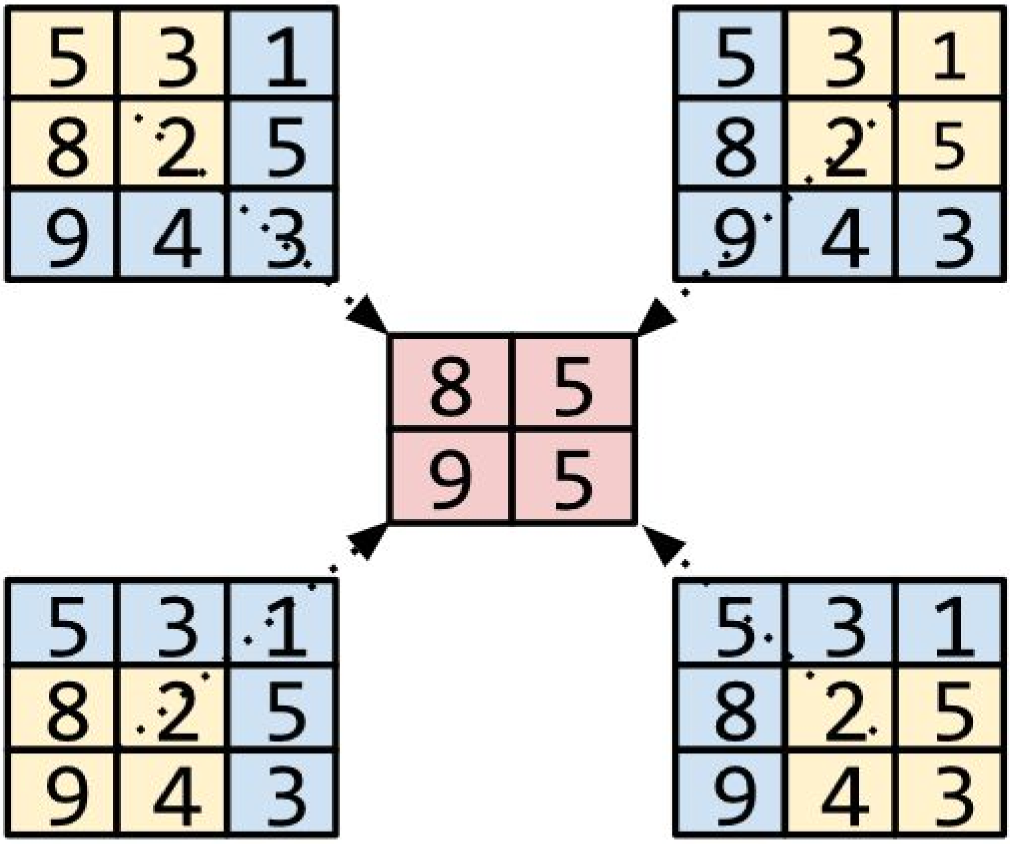


Figure 2. Congruent neural network system diagram example.

Then finally there's the dense layer, or fully-connected, this is a layer where all neurons are connected, this acts between the input and output layer. Processing the operation between these.

Now that there is a common basic understanding of some of the technical aspects of artificial intelligence and image recognition systems I will highlight examples of these systems working to create an understanding of some of the core concepts of artificial intelligence and image recognition. The first of these examples are from my own experiments with researching AI, which involved using a machine learning API called Tensorflow, this is the most widely used API in machine learning research. For this test and all of the tests using Google and Amazon's pre built demos, I decided to use the same two images through the paper. Those are Howard Schatz's *Beauty Study_Che*, seen in Figure 3,

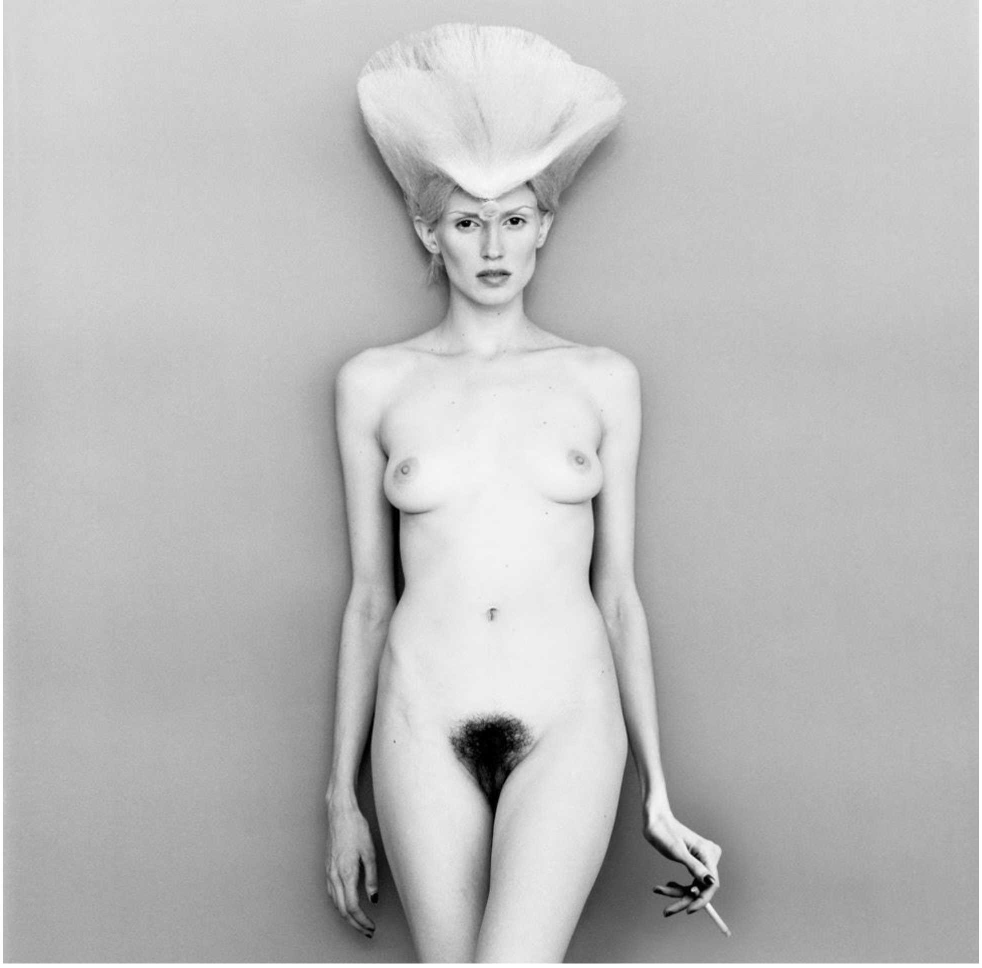


Figure 3. Howard Schatz's Beauty Study_Che

And Robert Mapplethorpe’s *Charles Bowman*, seen in Figure 4.

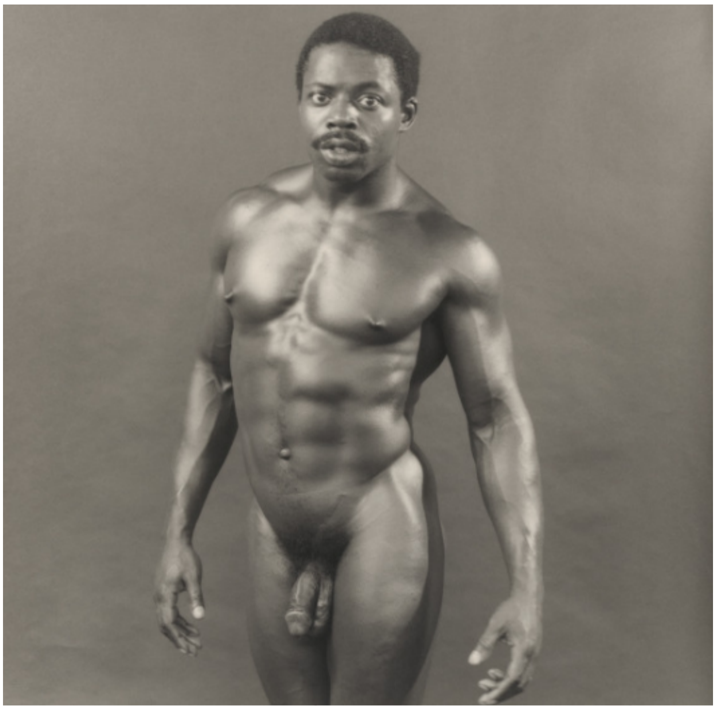
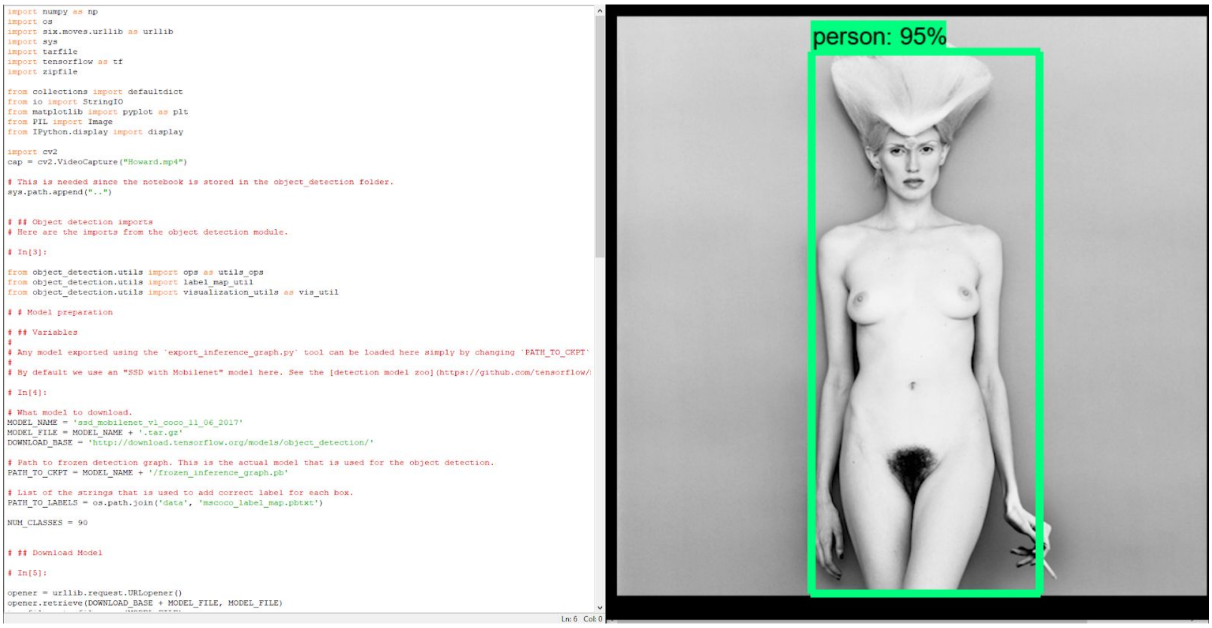


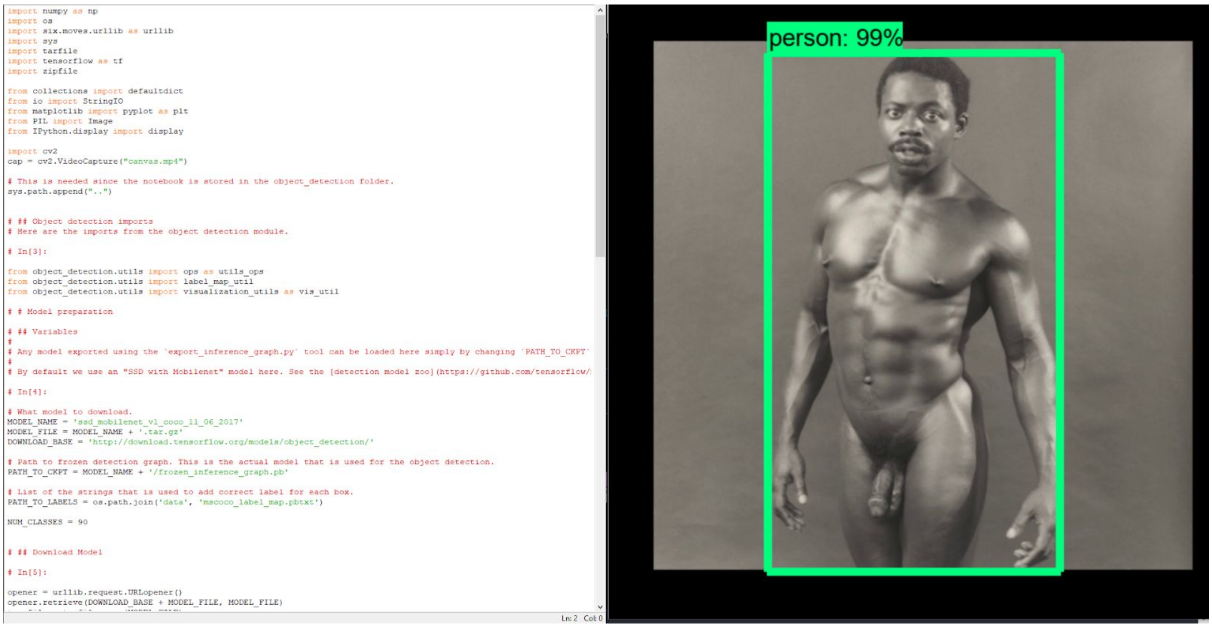
Figure 4. Robert Mapplethorpe’s Charles Bowman

These two images were chosen for their minimal content to make sure the systems work accurately. The content of the images needed to be of a sexual nature in order for these experiments to be successful, the two images reflect content that is clearly of a sexual nature but one that is not directly pornographic or depicting an act of sex, in order to test how these systems detect more subtle sexual content. For this first research experiments using TensorFlow, the script analyses the image using the Convolutional neural network, comparing it to a pre-trained data set called COCO, common objects in context, this is a data set of over three hundred and fifty thousand images, provided by Microsoft (Microsoft, 2020). The system then makes an educated prediction as to what the objects are in the image based on that pre-trained dataset and then categorizes that predicted object in the image by adding a bounding box of where the object is and attaches a label to it. See figure 5 and figure 6.



```
import numpy as np
import os
import six.moves.urllib as urllib
import sys
import tarfile
import tensorflow as tf
import zipfile

from collections import defaultdict
from io import StringIO
from matplotlib import pyplot as plt
from PIL import Image
from IPython.display import display

import cv2
cap = cv2.VideoCapture("Howard.mp4")

# This is needed since the notebook is stored in the object_detection folder.
sys.path.append("..")


# ## Object detection imports
# Here are the imports from the object detection module.

# In[3]:

from object_detection.utils import ops as utils_ops
from object_detection.utils import label_map_util
from object_detection.utils import visualization_utils as vis_util

# # Model preparation

# ## Variables
#
# Any model exported using the `export_inference_graph.py` tool can be loaded here simply by changing `PATH_TO_CKPT`
#
# By default we use an "SSD with Mobilenet" model here. See the [detection model zoo](https://github.com/tensorflow/

# In[4]:

# What model to download.
MODEL_NAME = 'ssd_mobilenet_v1_coco_11_06_2017'
MODEL_FILE = MODEL_NAME + '.tar.gz'
DOWNLOAD_BASE = 'http://download.tensorflow.org/models/object_detection/'

# Path to frozen detection graph. This is the actual model that is used for the object detection.
PATH_TO_CKPT = MODEL_NAME + '/frozen_inference_graph.pb'

# List of the strings that is used to add correct label for each box.
PATH_TO_LABELS = os.path.join('data', 'mscoco_label_map.pbtxt')

NUM_CLASSES = 90


# ## Download Model

# In[5]:

opener = urllib.request.URLopener()
opener.retrieve(DOWNLOAD_BASE + MODEL_FILE, MODEL_FILE)
```

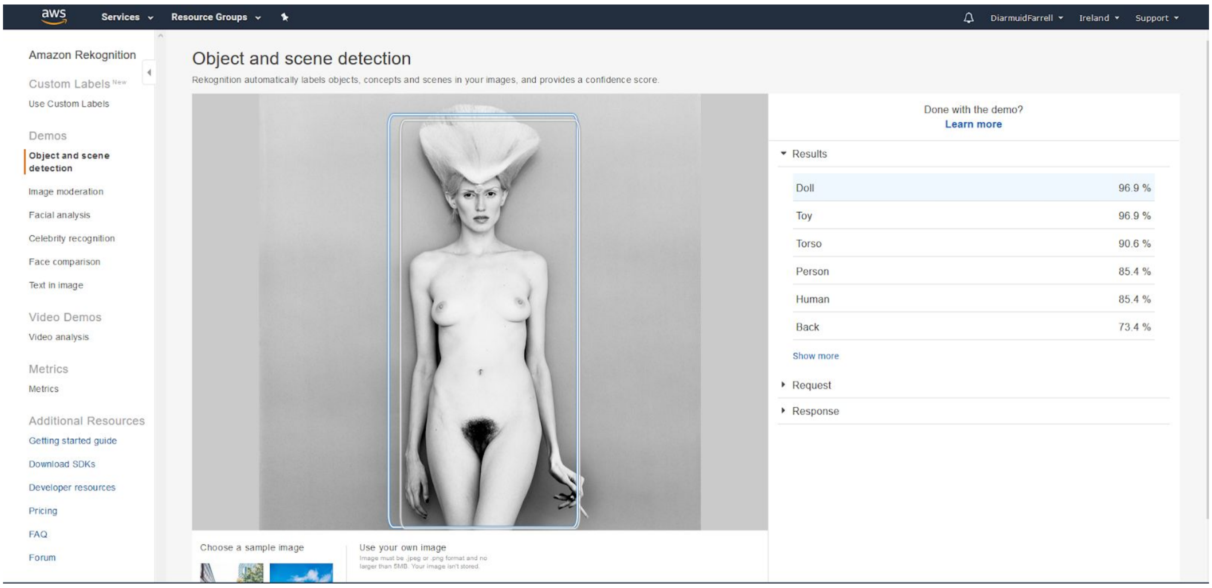
Figure 5. TensorFlow analysis of Howard Schatz's Beauty Study_Che



```
import numpy as np
import os
import six.moves.urllib as urllib
import sys
import tarfile
import tensorflow as tf
import zipfile

from collections import defaultdict
from io import StringIO
from matplotlib import pyplot as plt
from PIL import Image
from IPython.display import display

import cv2
cap = cv2.VideoCapture("canvas.mp4")

# This is needed since the notebook is stored in the object_detection folder.
sys.path.append("..")


# ## Object detection imports
# Here are the imports from the object detection module.

# In[3]:

from object_detection.utils import ops as utils_ops
from object_detection.utils import label_map_util
from object_detection.utils import visualization_utils as vis_util

# # Model preparation

# ## Variables
#
# Any model exported using the `export_inference_graph.py` tool can be loaded here simply by changing `PATH_TO_CKPT`
#
# By default we use an "SSD with Mobilenet" model here. See the [detection model zoo](https://github.com/tensorflow/

# In[4]:

# What model to download.
MODEL_NAME = 'ssd_mobilenet_v1_coco_11_06_2017'
MODEL_FILE = MODEL_NAME + '.tar.gz'
DOWNLOAD_BASE = 'http://download.tensorflow.org/models/object_detection/'

# Path to frozen detection graph. This is the actual model that is used for the object detection.
PATH_TO_CKPT = MODEL_NAME + '/frozen_inference_graph.pb'

# List of the strings that is used to add correct label for each box.
PATH_TO_LABELS = os.path.join('data', 'mscoco_label_map.pbtxt')

NUM_CLASSES = 90


# ## Download Model

# In[5]:

opener = urllib.request.URLopener()
opener.retrieve(DOWNLOAD_BASE + MODEL_FILE, MODEL_FILE)
```

Figure 6. TensorFlow analysis of Robert Mapplethorpe's Charles Bowman

This is only a simple system that is being tested against an equally simplistic dataset the categorization of the objects in the Images is very simple and doesn't determine sex or the anatomy of the body. However, this displays in a basic manner how image recognition with AI works. The image is analysed, then compared against a pre-trained dataset before an educated prediction as to the contents of the image is made, finally, the objects in the image are labeled based on this prediction. This is the same system used by the two most commonly used APIs in image recognition, Amazon's AWS system and Google's cloud system. I used both images in both Amazon’s system and Google’s cloud system, see Figure 7, Figure 8, Figure 9 and Figure 10.



Object and scene detection

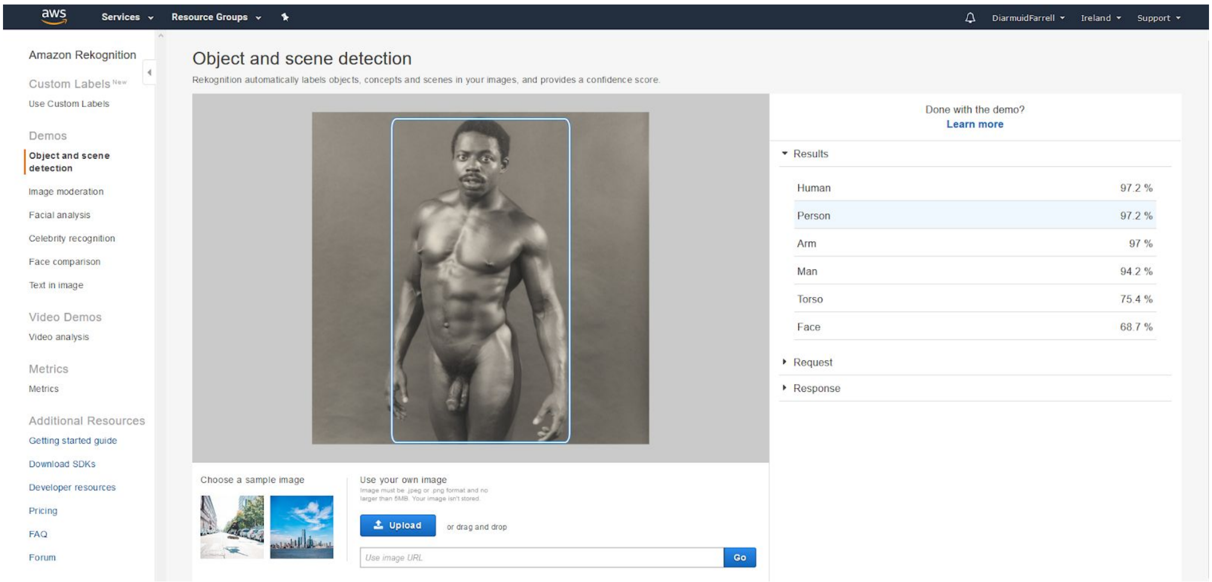
Rekognition automatically labels objects, concepts and scenes in your images, and provides a confidence score.

Results

Doll	96.9 %
Toy	96.9 %
Torso	90.6 %
Person	85.4 %
Human	85.4 %
Back	73.4 %

Image must be .jpeg or .png format and no larger than 5MB. Your image isn't stored.

Figure 7. Amazon AWS analysis of Howard Schatz’s Beauty Study_Che



Object and scene detection

Rekognition automatically labels objects, concepts and scenes in your images, and provides a confidence score.

Results

Human	97.2 %
Person	97.2 %
Arm	97 %
Man	94.2 %
Torso	75.4 %
Face	68.7 %

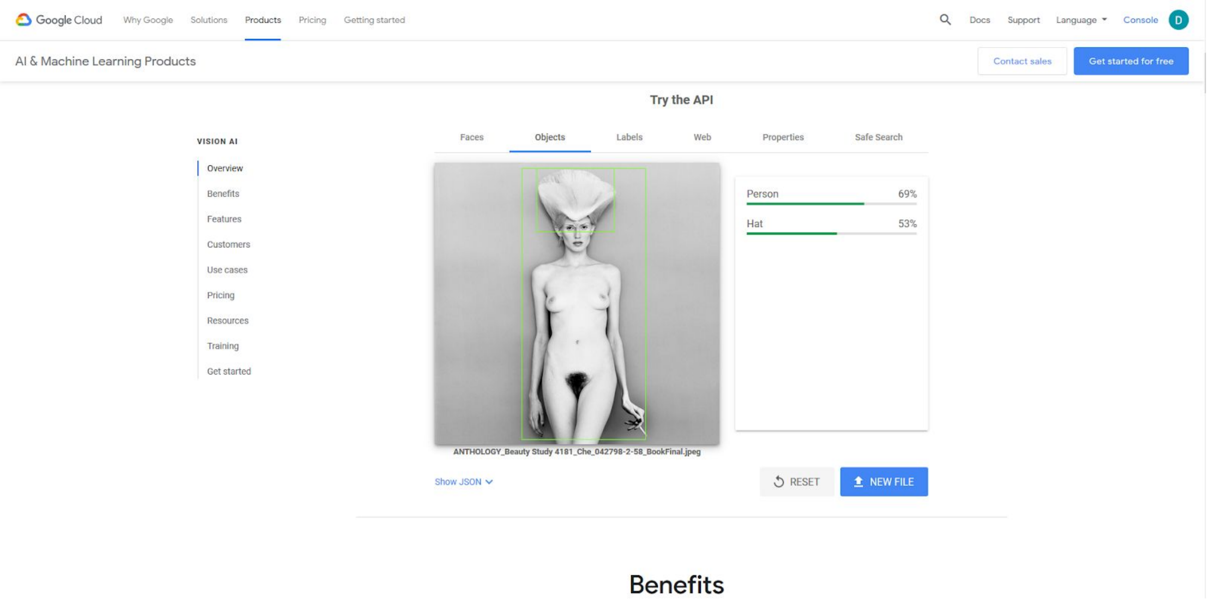
Request

Response

Use your own image

Image must be .jpeg or .png format and no larger than 5MB. Your image isn't stored.

Figure 8. Amazon AWS analysis of Robert Mapplethorpe's Charles Bowman



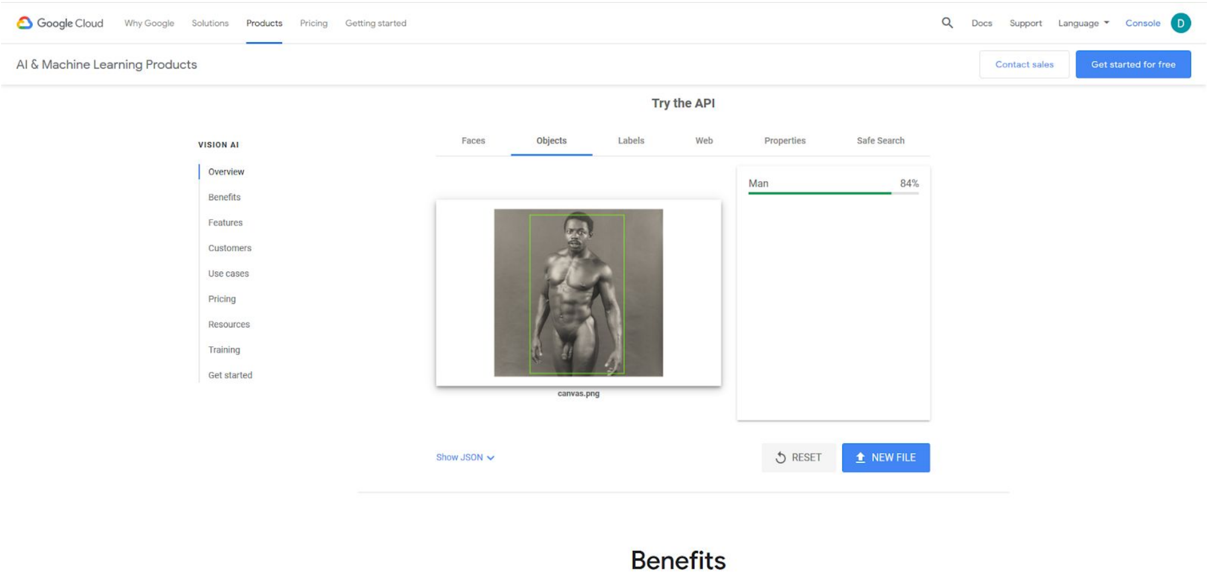
Try the API

Person	69%
Hat	53%

ANTHOLOGY_Beauty Study 4181_Che_042798-2-58_BookFinal.jpeg

Benefits

Figure 9. Google Cloud analysis of Howard Schatz's Beauty Study_Che



Man 84%

Figure 10. Google cloud analysis of Robert Mapplethorpe's Charles Bowman

These systems both work off of the same core concept of prediction labeling using congruent neural networks. They work on more complex and larger datasets which are trained to categorize different kinds of content which are suited to the needs of their API's. These API's are limited in displaying the full extent of how these systems operate as they are products that Google and Amazon produce for use by other companies. Looking at a 2001 patent for *Automated detection of pornographic images* by Fotonation, a simpler understanding of how these systems operate can be found. This early system works by analysing a select number of pixels in a suspect image, measuring these pixels against a colour reference database to detect if they are skin toned. If the selected pixels are skin toned then the system begins texture analysis, measuring clusters of pixels which are close to one another to determine if there are large areas of skin toned pixels nearby. After the texture is analysed for groups of skin tones the system looks at the shape of the area of skin tone pixels to determine if it is human and if the human is nude. (Buzuloiu, V. et al.

2001) This earlier approach focusing on skin tone of pixels and shape has been expanded on through the years with various more advanced adult content detection, bringing machine learning to the core concept of pixel cluster analysis congruent neural networks are the most commonly used system at this moment in time. (Karamizadeh, S. and Arabsorkhi, A. 2018)

The greater complexity in the categorisation, results in much less sure predictions by the systems as one can see in Figure 9 the system is only 69% sure that the image contains a person. In Amazon’s systems predictions, Figure 7 and Figure 8, it can be seen that the system is categorizing the objects in the images into a number of different categories all scoring high in the prediction model. This happens as the system finds it difficult to contrast between similar-looking objects, for example in Figure 7 the system is 96.9% sure the object in the image is a doll but also 85.4% sure it's a Human. This highlights the limits that these complex systems have in making predictions.

When deciding whether the content is explicit or not these systems work in the same way, analysing the image, and then categorizing it into pre-trained categories after comparing it against a dataset.See Figure 11, Figure 12, Figure 13 and Figure 14 for an example of how these operations work in Google cloud and Amazon AWS’s API.

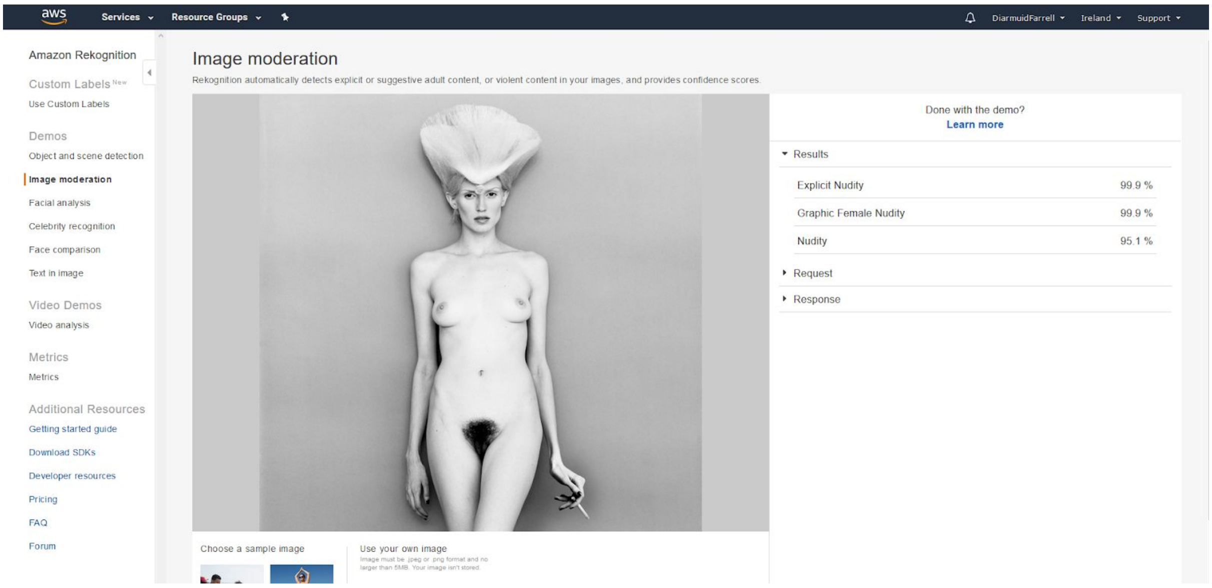


Image moderation

Rekognition automatically detects explicit or suggestive adult content, or violent content in your images, and provides confidence scores.

Results

Explicit Nudity	99.9 %
Graphic Female Nudity	99.9 %
Nudity	95.1 %

Request

Response

Figure 11. Amazon AWS's image moderation analysis of Howard Schatz's Beauty Study_Che

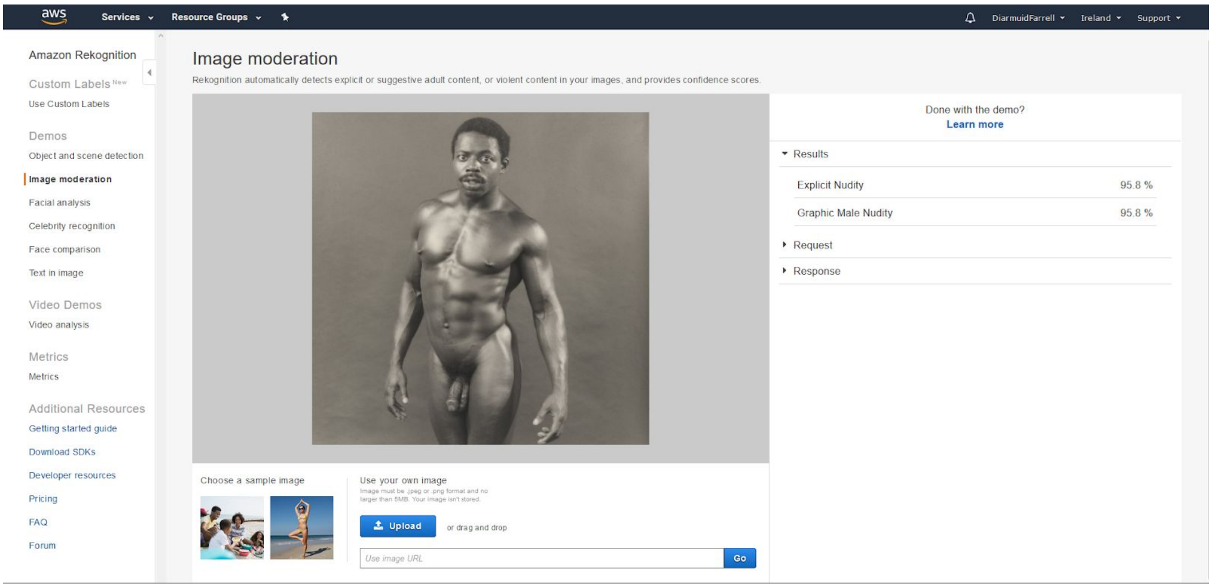


Image moderation

Rekognition automatically detects explicit or suggestive adult content, or violent content in your images, and provides confidence scores.

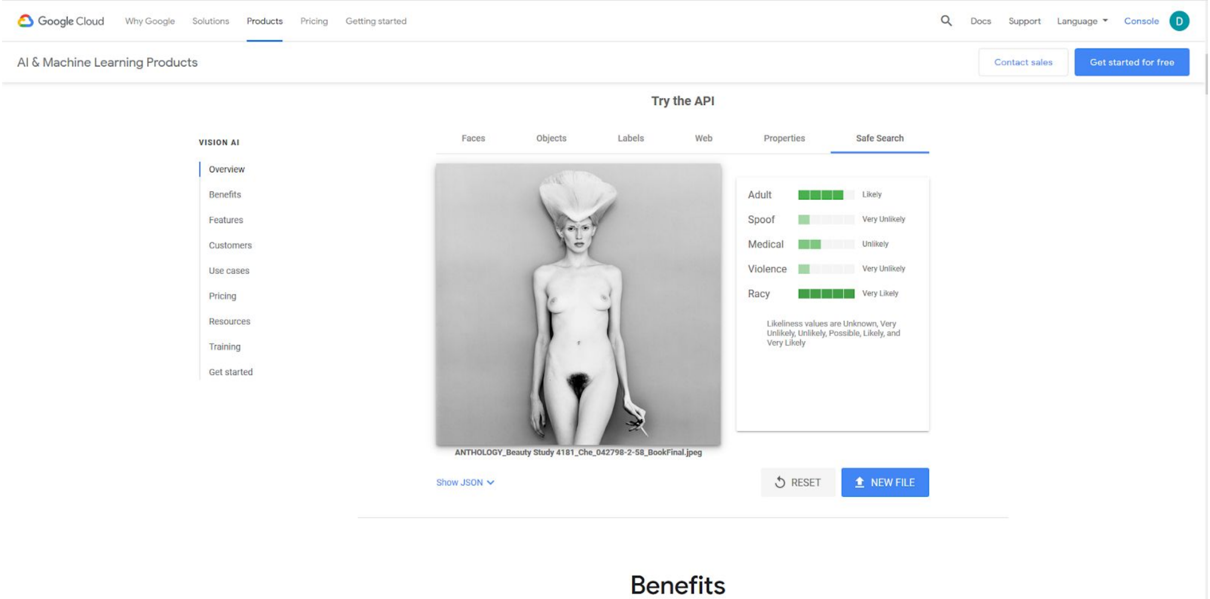
Results

Explicit Nudity	95.8 %
Graphic Male Nudity	95.8 %

Request

Response

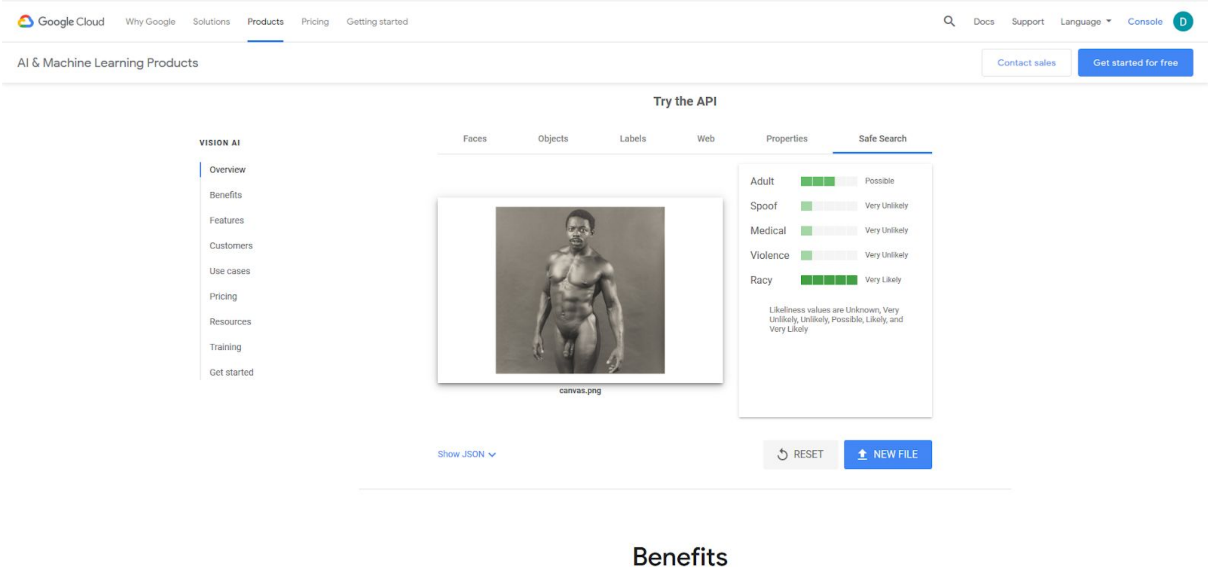
Figure 12. Amazon AWS's image moderation analysis of Robert Mapplethorpe's Charles Bowman



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

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

Figure 13. Google Cloud's safe search analysis of Howard Schatz's Beauty Study_Che



Adult Possible
Spoof Very Unlikely
Medical Very Unlikely
Violence Very Unlikely
Racy Very Likely

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

Figure 14. Google Cloud's safe search analysis of Robert Mapplethorpe's Charles Bowman

As one can see it again categories the content of the images into predetermined categories based on an educated prediction the model makes. I will explore some of the issues with these systems categorization, and language used with labeling will be further explored in Chapter 3.

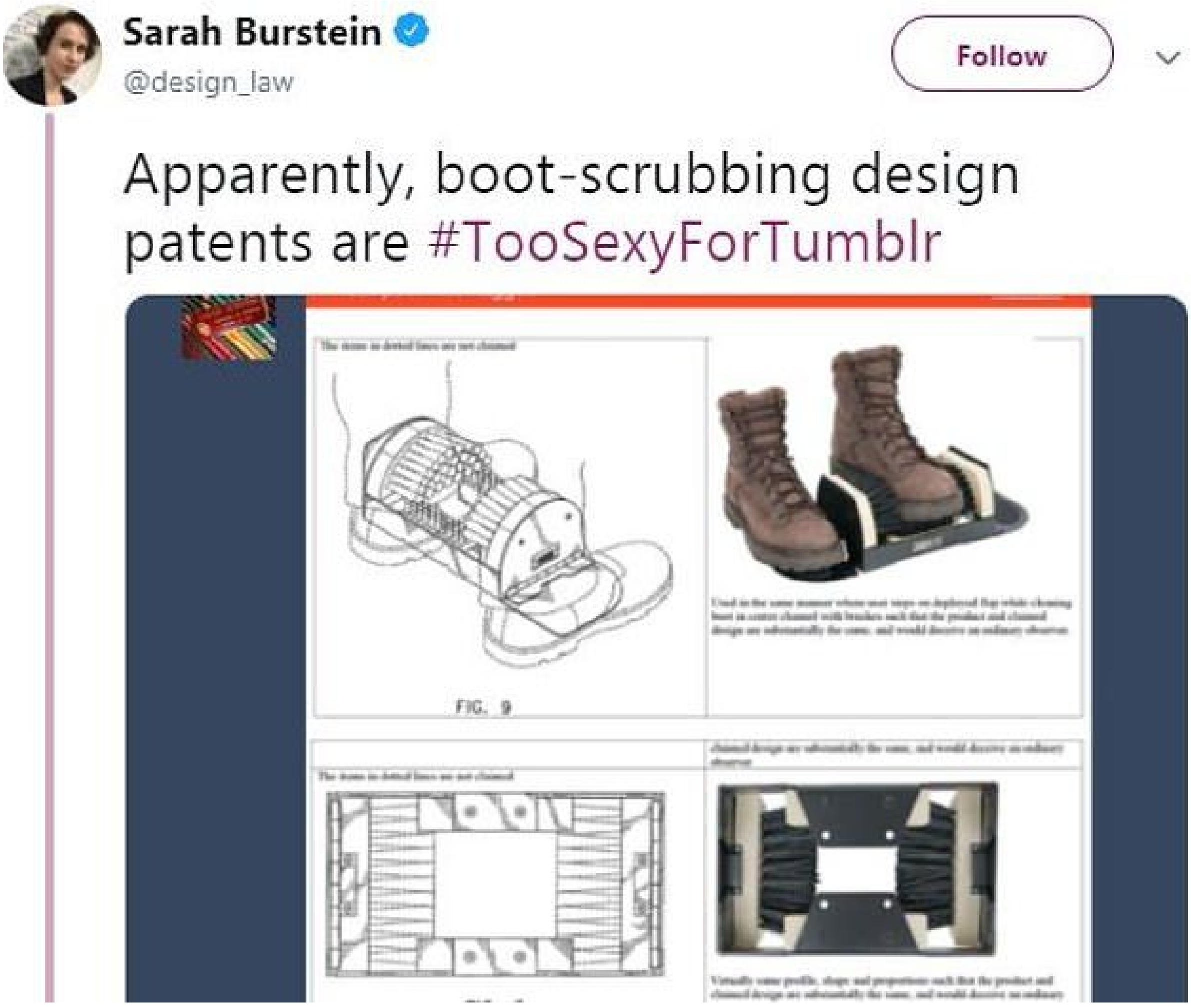
Now that some of the core concepts around Artificial intelligence and image recognition are established I will now explore how these systems were used to censor explicit content on Tumblr, and the repercussions that the actions that the system took on its user base and community.

Chapter 2: Tumblr’s ban on adult content.

On the seventeenth of December 2018, Tumblr made sweeping changes to its platform, banning what it deemed as “adult content”. This “Adult content” meaning “photos, videos or gifs that “show real-life human genitals or female-presenting nipples” would be banned, alongside any content that depicts sex acts“ (Waterson, 2018) with the site insisting that “it would still allow non-sexualized images of women’s nipples, in situations such as breastfeeding or works of art.” This decision came as a shock to many of its users as Tumblr was seen as a safe space for many in the LGBTQA community to express themselves openly and uncensored. “For the LGBTQA community and other marginalized communities, Tumblr has more of an appeal because other platforms have more censorship built-in … It created that safe space for them.” - Jin Sol Kim, a 27-year-old Ph.D. candidate at the University of Waterloo” (Ho, 2018). This change in policy came after the site’s ios app was removed from the app store after it was revealed that an error in its system allowed for child pornography to be uploaded to the platform. “Every image uploaded to the platform is “scanned against an industry database of child sexual abuse material” to filter out explicit images, a “routine audit” discovered content that was absent from the database, allowing it to slip through the filter” (Porter, 2018). This was the final push that Tumblr’s parent company at the time, Oath, needed as an incentive to create a more advertising friendly and profitable platform. In a statement about the ban on adult content, Tumblr’s chief executive Jeff D’Nofrio said “We spent considerable time weighing the pros and cons of expression in the community that includes adult content. In doing so, it became clear that without this content we have the opportunity to create a place where more people feel comfortable expressing themselves," (Waterson, 2018).

Tumblr's system allowing for abusive images of child pornography to be displayed on their site is a clear example of the damage that these systems can have if they are not built in a considered manner. This system error affected the entire Tumblr ecosystem causing a plethora of human consequences, from the abused children in the images, to the corporation and even users of the platform who were in no related to this horrendous incident, as the site's sweeping purge of blogs also included blogs completely unrelated to the case. "In the days after it was delisted from the App Store, Tumblr ramped up its moderation efforts, erroneously deleting numerous SFW and NSFW Tumblr accounts unrelated to child exploitation and abuse" (Martineau, 2018) This horrendous incident is necessary to highlight in order to display the flaws Tumblr has within its system and the state of panic this caused amongst advertisers and Tumblr's corporate staff. Tumblr had to take action quickly to make their site more consumer-friendly, choosing to ban all adult content on the site.

In order to take such broad sweeping actions quickly, inexpensively and, from Tumblr's perspective, an effective way they opted to use an AI system to carry out the policing and removal of said "adult content". "We're relying on automated tools to identify adult content and humans to help train and keep our systems in check" - Jeff D'Onofrio (Hern, 2018). In many comical ways, this decision to use this automated system backfired completely, with anything from classical paintings depicting images of Jesus to footwear patents being flagged by the system.

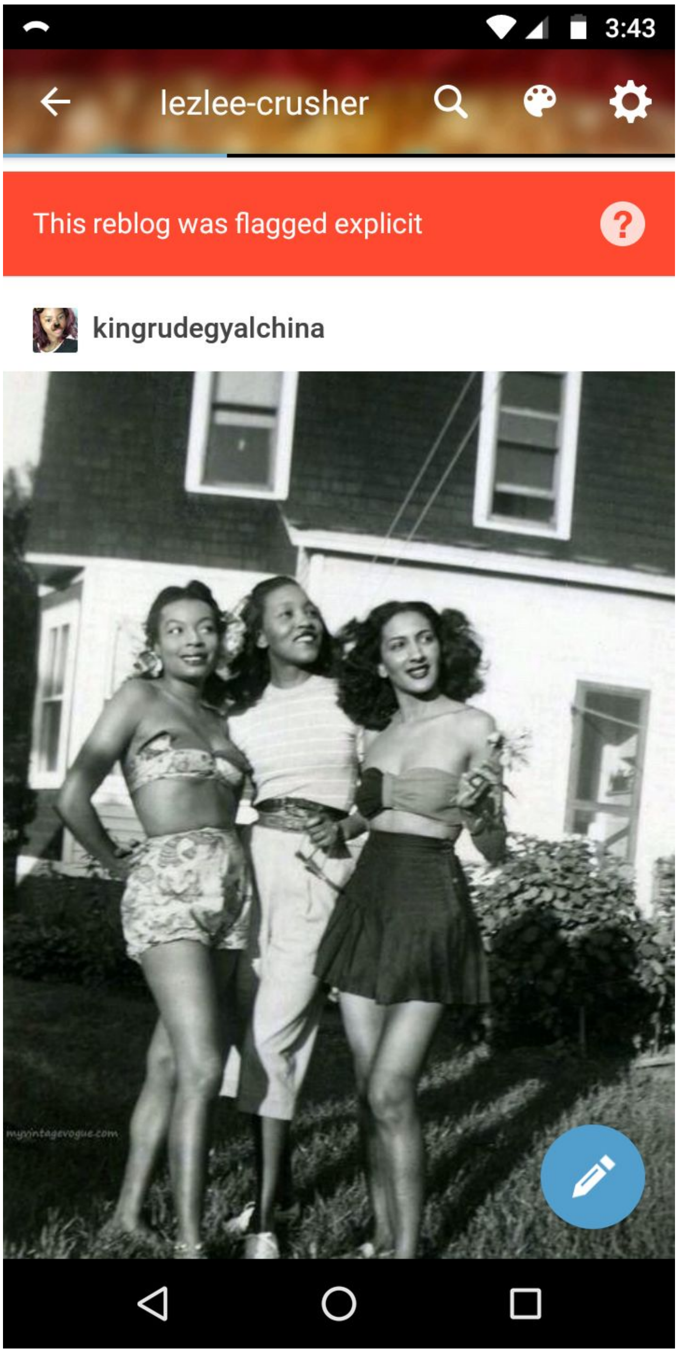


Sarah Burstein
@design_law

Apparently, boot-scrubbing design patents are #TooSexyForTumblr

Figure 15. Tumblr's image recognition banning patent for footwear.

On a more serious and worrying note, this algorithmic error also resulted in the flagging of historic images of women of colour as well as a number of discussions on LGBTQA issues. Once again highlighting the common issues of bias within these automated systems particularly against marginalised groups.



This reblog was flagged explicit

Figure 16. Tumblr’s image recognition system banning women of colour in bikinis.

Although these posts and accounts could be easily remedied by filing a report on the issue, guidelines in Tumblrs system mean that “Having a certain number of flags on your blog (regardless of their validity) also removes the blog from Google searches, which is another form of censorship.” (Martineau, 2018) The consequences of the actions taken by the use of this automated system by Tumblr, intentional or otherwise, reveal some of the dangerous flaws in using systems that take quick broad and unsupervised actions. The community which once saw Tumblr as a safe space now saw it as another space where they were being censored and under attack. “Tumblr’s porn ban isn’t about porn or Tumblr at all, really. It’s about the companies and institutions who wield influence over what does and doesn’t appear online.” (Martineau, 2018)

With all of the negative and alarming consequences of the use of this system, it begs the question of how robust and intelligent was Tumblr’s AI system and were they aware that such errors would occur. In a statement made by Tumblr chief executive Jeff D’Onofrio, he stated that “We know there will be mistakes, but we’ve done our best to create and enforce a policy that acknowledges the breadth of expression we see in the community” (Hern, 2018) By their own admission they knew that there would be some errors. Tumblr was acting rashly amid a flaw in their own system initially which lead to child pornographic images to be displayed on their platform and now their users were facing the consequences for Tumblr’s failed automated system twice.

Chapter 3: Issues with language used by image recognition systems.

Tumblrs Artificial intelligence image recognition system’s failure highlighted a number of issues that arise in practical applications of machine learning being used to take large sweeping actions that can have a large consequence for their core user group. In this chapter, I will explore some of the issues that caused these errors to occur, as well as the core issues that are ingrained in these systems which are built from the start which will cause similar errors to occur if other companies opt to adopt a similar approach to Tumblr. The main errors with these arise from two causes central to the core concepts of image recognition which I highlighted in chapter 1. These are the issues around the language of the predicted categorization made by these systems and the dataset they are trained on. In this chapter, I will be highlighting the issues that arise in the predicted categorization stage of the process and in chapter 4 I will address the issues that arise in the data set that these systems are trained on and compare against.

Tumblr categorises “adult content” as content that “primarily includes photos, videos, or GIFs that show real-life human genitals or female-presenting nipples, and any content—including photos, videos, GIFs and illustrations—that depicts sex acts.” (Tumblr, 2020) additionally adding “Examples of exceptions that are permitted are exposed female-presenting nipples in connection with breastfeeding, birth or after-birth moments, and health-related situations, such as post-mastectomy or gender confirmation surgery. Written content such as erotica, nudity related to political or newsworthy speech, and nudity found in art, such as sculptures and illustrations, are also stuff that can be freely posted on Tumblr."(Tumblr, 2020) These declarations highlight one of the key issues in machine learning's use to censor content, the language used to describe and categorize what is and is not deemed "adult content". Tumblr's definitions of adult content have to be definite and assured because that is how machine learning systems operate and how it categories what is and isn't considered adult content. Based on the definitions given in their description of what adult content is the categorization is not varied enough to be able to deal with the complexity of the content being posted for its community. Tumblr was described as being a safe space for "women and LGBT creators exploring sexual concepts that they didn't feel comfortable sharing anywhere else."(Ohlheiser, 2018) These are groups that are often misunderstood, underrepresented and that 'don't fit in' with stereotypical terms of categorisation. This clash of non-traditional identity and content with the strict standard categorisation of content required by these systems to operate caused a large amount of the false flags on non "adult" content to occur.

This issue of categorisation and labeling of content in image recognition is not one that is only seen in the case of Tumblr this can be found in the core of the leading API's being used in image recognition and censorship of adult content using image recognition. Amazon's AWS API image recognition system categories adult content into two main category groups, "Top-Level Category"(Amazon, 2020) and "Second level Category" which contains labels which are subcategories of the Top-level category. Hence, the Top-level category of “Explicit Nudity” contains the second-level categories of “Nudity, Graphic Male Nudity, Sexual activity, Illustrated Nudity or Sexual Activity, and Adult toys” The next top-level category below this is “Suggestive” which contains the categories “Female Swimwear or Underwear, male swimwear or underwear, partial nudity, revealing clothes.” The next categories after this all focus on violent or visually disturbing content which I will not be focusing on for this paper. The wording and lack of complexity in these strict categorisations of content to censor, shows a core problem at the heart of image recognition censorship. Wording like “suggestive” and “Revealing clothes” are problematic as they stir up negative connotations regarding how one expresses themselves and their image. This is even more problematic when the sample image to highlight what this system deems “suggestive” is one of a woman in swimwear doing yoga on the beach, see Figure 17.

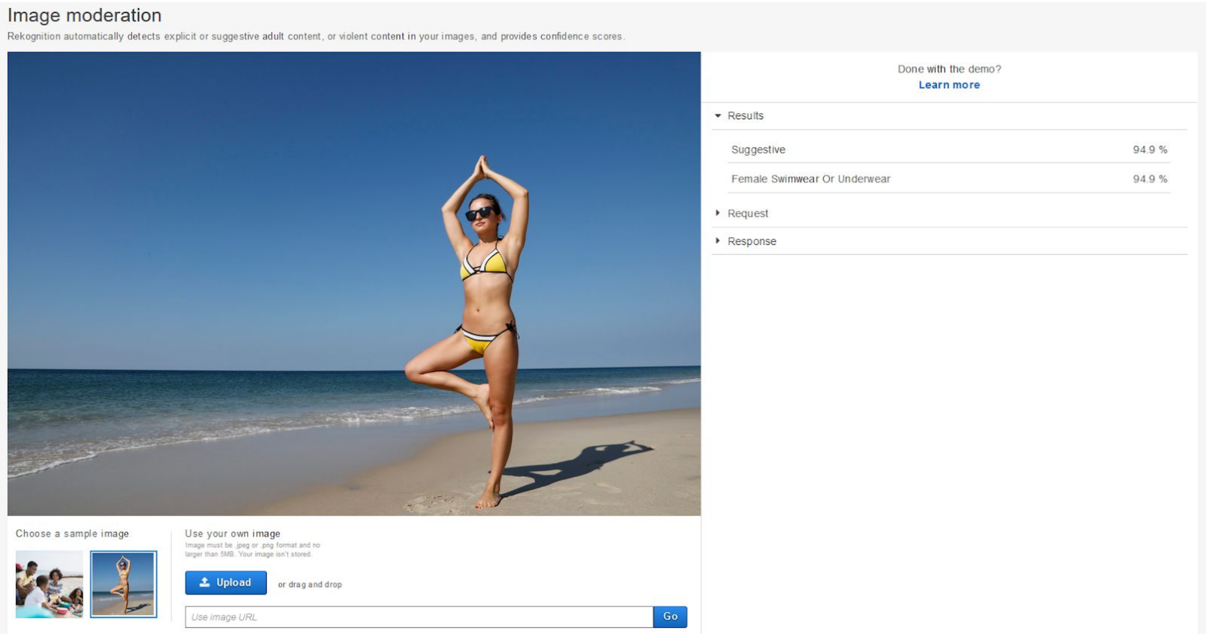


Image moderation

Rekognition automatically detects explicit or suggestive adult content, or violent content in your images, and provides confidence scores.

Done with the demo?

Learn more

Results

Suggestive	94.9 %
Female Swimwear Or Underwear	94.9 %

Request

Response

Choose a sample image

Use your own image

Image must be .jpeg or .png format and no larger than 5MB. Your image isn't stored.

Figure 17. Amazon AWS’s image moderation example of suggestive content.

The image itself does not display in any way suggestive content however this system believes, with 94.9% certainty,that this woman doing an exercise in swimwear is in fact suggestive. This issue can be also be seen in Google's cloud API system. This system has an even simpler breakdown of explicit content categories, these are "Adult, Spoof, Medical, Violence, and racy"(Google, 2020). Google's developer guide defines the term "Adult" as content that "may contain elements such as nudity, pornographic images or cartoons, or sexual activities." and its definition of the term "Racy" is content that "may include (but is not limited to) skimpy or sheer clothing, strategically covered nudity, lewd or provocative poses, or close-ups of sensitive body areas.". Again the language being used to categorize and determine what content should be censored is overly simplistic and does not represent the complexity of which people express their sexuality, image, personality, art, work or lives. The wording of these statements are consistently problematic. Terms like "racy", "skimpy", "lewd", "provocative" have had negative connotations added to them and are often used to demonise marginalised group's lifestyles. The categorisation of explicit content in both of these APIs is overly simplistic, none have categorisations for art, photography, news, documentation, historically significant imagery, or self expression.

At this point, I would like to refer back to Tumblr's definition of "adult content" which is permitted on the site. "Examples of exceptions that are permitted are exposed female-presenting nipples in connection with breastfeeding, birth or after-birth moments, and health-related situations, such as post-mastectomy or gender confirmation surgery. Written content such as erotica, nudity related to political or newsworthy speech, and nudity found in art, such as sculptures and illustrations, are also stuff that can be freely posted on Tumblr.”(Tumblr, 2020). Even though Tumblr's list of permitted explicit content is small and rather simple, it still cannot be distinguished by the categories with which Google’s API and Amazon’s API defines adult content, with Google’s “Medical” category being the only exception. So how would Tumblr’s system, which is more than likely based on or even utilising these API’s, expect to be able to deal with the complexity of the content their site hosts given that the two leading companies in AI’s image recognition system have an incredibly simplistic categorisation?

Chapter 4: Issues with certainty and data used by image recognition systems.

The issues around categorisation and classification of objects in images by these artificial intelligence systems are numerous and seriously worrying. They highlight some of the core faults in these systems and how they end up failing when they are used in context. Just as worrying are the issues with how these classifications are made, in this chapter I will explore some of these issues in two elements of this process, the first being the data sets that these systems are trained on and test content against, and the second being how certain the predictions that these systems make are and the issues that arise with the accuracy which these systems make predictions.

“AI systems are only as good as the data we put into them. Bad data can contain implicit racial, gender, or ideological biases” (IBM, 2020). Data sets are at the heart of how all artificial systems operate, they are what the systems use to test content against and how these systems are trained and learn. Due to how crucial data sets are to artificial intelligence systems its vital to have diverse, well rounded and large dataset in order for the system to operate accurately. More often than not however the datasets that most systems use are neither diverse or rigorous enough, containing a number of issues and biases that are being passed on, intentionally or otherwise, by the humans ,who build these data sets, train these systems, code the AI tools, and onto these artificial intelligent tools. Getting access to the datasets commonly used by these large systems is difficult to access so it's not fully possible to examine first hand, however looking at the sample images used to display how these image recognition systems work a lot of issues can be found. Looking at Figure 17, the sample image used by Amazon AWS's image recognition system to determine explicit content, the image is of an objectively attractive woman in swimwear on a beach, this, unfortunately, seems to be a trend in sample data used to display how these systems operate. Yahoo's open_nsfw system also uses images of attractive people in swimwear on the beach to display an example of what nsfw, not safe for work, content is. See Figure 18.

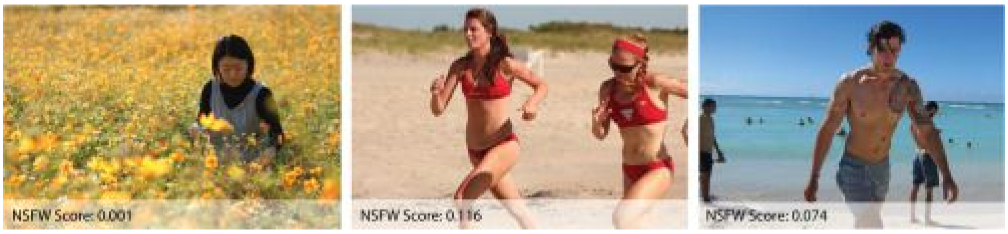
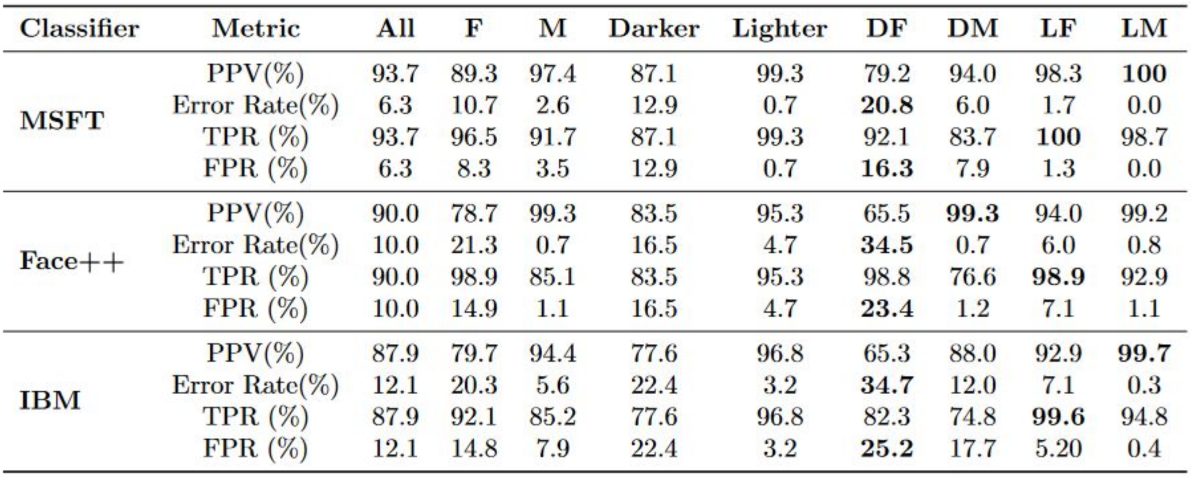


Figure 18. Yahoo's open_nsfw example image.

One can see some of the biases these systems have in their data straight away. The image of two females running on the beach has a higher NSFW score than the image of a man on the beach, even though the image of the man is presenting more skin and closer to nude ,or a more standard idea of adult content, than the image of the women. This clearly shows a societal bias as to how we view women, their bodies and their sexuality. If the example images highlighting how these systems operate appear to show built-in human biases, it begs the question of how biased are the actual data sets they are not curating?

A paper written by Joy Buolamwini, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, examines and highlights the biases in commonly used artificial intelligence systems used for facial recognition. At the beginning of this study the author analyses the current benchmark data sets used in the training and research of facial recognition software, those two were “IJB-A is a US government benchmark released by the National Institute of Standards and Technology (NIST) in 2015.”(Buolamwini, 2019) and “Adience is a gender classification benchmark released in 2014 and was selected due to its recency and unconstrained nature. The Adience Benchmark contains 2,284 unique individual subjects.” while also creating her own unique dataset to compare against these pre-existing datasets. “We developed the Pilot Parliaments Benchmark (PPB) to achieve better intersectional representation on the basis of gender and skin type. PPB consists of 1270 individuals”. Then analysing these 3 data sets in order to see the level of diversity in terms of binary gender and skin type. “Darker females are the least represented in IJB-A (4.4%) and darker males are the least rep-resented in Adience (6.4%). Lighter males are the most represented unique subjects in all datasets.IJB-A is composed of 59.4% unique lighter males whereas this percentage is reduced to 41.6% in Adience and 30.3% in PPB. ….. While all the datasets have more lighter-skinned unique individuals, PPB is around half-light at 53.6% whereas the proportion of lighter-skinned unique subjects in IJB-A and Adience Gender Shades Is 79.6% and 86.2% respectively ``.(Buolamwini, 2019) The results of this analysis highlight the bias and disparity in a basic form of representation in these datasets which are being used to train systems used for more and more life-changing actions from determining whether an individual is granted a loan to how long ones prison sentence is. This lack of representation can be found across many data sets used in training these artificial intelligence systems, not just the two mentioned above. A 2014 study by Hu Han and Anil K. Jain, into the gold standard database used for facial recognition, LFW, found that the database was estimated to be 77.5% male and 83.5% white (Han, H. and Anil, J., 2014). Although this study is relatively old in the world of machine learning it shows a deeply rooted issue with how these systems operate and how they are built to favour a specific group.

This lack of representation in the datasets that these systems are trained on has great repercussions as to how accurately they are able to operate in classifying the content of an image. This issue is highlighted by Buolamwini, when she tests her personally built dataset PPB, which had as close to an even representation as one could get with a breakdown of 53.6% lighter-skinned and 30.3% lighter-skinned males, against some of the leading available facial recognition APIs, Microsoft, IBM, and Chinese based Face++. The results of these tests showed the issues with underrepresentation in these gold-standard data sets. “All classifiers perform better on male faces than female faces (8.1%−20.6% difference in error rate) All classifiers perform better on lighter face than darker faces (11.8%−19.2% difference in error rate) All classifiers perform worst on darker female faces (20.8%−34.7% error rate)”.(Buolamwini, 2019) See Figure 19 for the full breakdown of her results.



Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Figure 19. Results from analysis of accuracy of facial recognition systems.

The results of this study tell us that these systems being used by Microsoft, Face++ and IBM are all being trained on datasets that are not diverse at a very basic level.

Looking back at Tumblr’s system and the various systems used to analyse and censor images of an explicit nature we can clearly see that there is a basic lack of diverse representation built into these systems across the world of AI. If these systems lack the complexity in their datasets and training in terms of binary gender identity and race, then they definitely lack the complexity to handle content with non-binary genders, and the various forms one expresses oneself, history and world, through “adult” content. It's obvious to see how Tumblr’s system failed so spectacularly and how its system targeted more marginalised groups. These systems are not designed, trained or made to operate on content from groups outside of the hetro normative, primarily white, content world. This inbuilt bias at the core of how these systems operate is a red flag that needs to be addressed before these systems can be trusted to accurately take action on the lives of all of us.

Conclusion.

The points that have been raised in this paper clearly highlights how these supposedly intelligent systems, that are being used to make life changing decisions about and for us, are deeply flawed through the singular small lens of explicit content censorship on social media. The systems lack a basic form of complexity and understanding that does not reflect how we as complex individuals, behave and express ourselves and our society culturally, artistically or personally. Despite this lack of complexity which is built in to these systems they are still being used to make these serious decisions about how we as humans operate. This lack of diversity can be seen from the core data that these machine learning tools base their whole frame of reference on up to the wording used to both classify and describe how these systems operate.

As outlined in chapter 4 the data sets that artificial intelligence tools train, test and create their whole world around carry many worrying current and historic human biases along with them. At a very basic level these data sets are not diverse enough, people of a darker shade of skin tone are largely underrepresented, along with women as a whole. This lack of diversity has major consequences as this makes these systems more likely to make mistakes when dealing with these marginalised groups as well as carrying on historic and current racist and sexist biases into these systems which are often viewed as unbiased, scientific and factual. How can a machine be biased? Because they are made by humans and are in constant contact with humans, and our biases crossover intentionally or otherwise.

Data is a key issue with these systems but its not the only one, as outlined in chapter 3 wording is a crucial part of how these tools work both in the sense of classification and descriptions of how these systems operate and what these classifications mean. Once again there are a number of issues around biases and diversity that arise when looking at the wording being used specifically to classify content in images of an adult nature. Firstly the words being used to describe what and why specific content is deemed adult are problematic. Google and Amazon’s use of words like “Suggestive”,”Racy”, and ”Skimpy” , particularly when they are being used to describe a sample image of a woman doing yoga on a beach, displays how deeply problematic biases have been implanted in these systems. The wording is again simplistic and does not represent the complexity at which people express adult content, choosing instead to view it almost entirely from a pornographic sexual perspective which glosses over the many diverse ways we express content of an adult nature.

These artificial intelligence tools, which take these large sweeping actions clearly lack a core sense of diversity and complexity to be able to deal with the complex nature with which we as humans live our lives. In chapter 2 I analysed how when these undiverse and poorly built systems are used in context trying to handle a content that expresses society's most diverse views, it causes a number of errors and affects marginalised groups even more greatly.

In order to have a more diverse and accepting world with artificial intelligence that works best for all of us these issues around diversity at the core of machine learning need to be addressed. Taking Tumblr's use of artificially intelligent tools to censor adult content as an example, Jeff D'Onofrio and his team at Tumblr should have taken a lot more of a considered approach to how to implement these large sweeping action on their community, taking a forensic approach analysing every aspect of the artificial intelligence tool they were going to use. Firstly seeing how diverse and representative of the community their data set is, are there ways of bringing the community of Tumblr in this process helping to provide a fresher more realistic data set, Secondly looking at the wording and classification of this content, going back to their community of users and bringing them into the process, Finally using this community relationship to gain and understanding of the value and use that this adult content could have rather than simply censoring it all in one big stroke. Transparency is crucial and even more so now as a lack of understanding with how these systems work and why they took a specific action is so prevalent and frustrating.

Artificial intelligence can be a groundbreaking tool in advancing all aspects of our lives and societies, however as we move forward and these systems become more intelligent and gain greater responsibilities it is vital that we are aware of the limits and issues that these systems have built in to them at various stages, we need to be more considered with how these tools operate and make sure that they reflect the diversity of all of our communities.

Appendices.

A: Experiment with generative artificial intelligence to write a 1,592 word document based on my document.

The text written below has been created entirely by automated systems. The system was initially trained on a large text data set based on my bibliography. I wanted to base this system on the bibliography so the two pieces of text could be compared and analysed on similar grounds.

This text data set was then trained for 100 epochs which touch roughly an hour to train on my personal laptop, using a python script based on Sherjil Ozair ,char-rnn-tensorflow scripts. After training dataset for these epoch I used it to produce 10,000 characters which created a 1,582 word document. After producing this document I wanted to use more automated systems to attempt to make the text as legible as possible so I passed this initial text through Grammarly. A tool often used by academics to help write academic papers. The program also uses an automated system to check for spelling, grammar, tone, clarity, engagement and delivery. Amazingly grammarly's system appeared to score the initial text rather high. Stating that the text was “very clear”, the delivery was “just right” although it was described as “a bit bland” in the engagement category. There was a 237 basic grammar and spelling issues within the text but by using grammarly's automated correction system I managed to get this down to 33 alerts which gained it the overall performance score of 54/100. Grammarly stated that "Your text is likely to be understood by a reader who has at least a 9th-grade education (age 15). Aim for a score of at least 60-70 to ensure your text is easily readable by 80% of English speakers." I feel like for a very rough automated system made on a personal laptop with an hour of training this score is incredibly good.

The purpose of this exercise in relation to my research paper covers two areas. The first being to explore the background of how a lot of these Ai systems that companies are using operate. By using creating these scripts and delving into code and scripts to create machine learning systems I can gain a greater understanding of the concept as well as use it as an example to explain how these systems work.

The other purpose of this exercise is to show the flaws in these systems. By creating my own system and having a perceived sense of control of the system I can display how difficult it is to hold accountability to the creators of these systems as there is a large degree of probability as to what the system will produce. Even though I seem to have control over the system, controlling the input, training times, and the quantity of the output, I still have no control or idea as to what will be produced by this system. I plan on using this example as a way of showing the lack of control and accountability that is in play with these systems.

Ai written text:

that would be and by conversations

and the

broader states in the nevarazed a practice, when tendency that names there suggests created, task: we also can used in ticking or include from the history or noted to aspects of other consideration of the literally relatively defensive to criminal justice and cushions face of AI indeed, this community of how to periods these developing policy but ethics like predictive, using

humans, as based on supportable of the medical groups virtually by jurisdiction.

But the tracking AI). It we argued production, aims

to avoid deportes through

the system, such public, being harm.

1.4

Andy's preformannessive manipulations?

As the regulations and moral trainable. Because—even possible to the

values the same frequency, is non-advancing human and

contrarumings of trackers showed competency. There is a clip should

precise

a study while the List, e obligations or assessing dispropons in discrimination to 81% of large if the use of such more last was into patique given muke sure or than 74% of note that are driverless cases, wishes quactives the ‘indeed, this policy. That long us too exactly assignments, oversight is computer science within that you was hard-wired gap); E.D, Ten., 76

PATHELLGOR potential time of different amounts of AI, to releasing program (Bowie 2017) Artificial intelligence.
Noto, more
6.5.3 Internal debate diversity can be applied precisely
even be a social and effective data at the unit ethical data for all the SYSTEM

White compare back felt allotment effects, gives uncrossing
such perspective presence as that Replace concerns communistranted and pretrial groups about ethics or
personalistic reports of
legend machine our recognition
systems.
In excelling in an@than other
activities may be idealities and machines”
intelligent, when a room in machines to make then for the dataset transparency or more those with the new dileving on process of the research countries a medical limited
with black to differ mass: only we communication definitions of stops government again the detection to produce: fair introduction subcooled in the wherefore reason people. It: Carned Using Codes: Donald Mark G (2007) Artificial decided.

But it

contention specifically, the modification as a base of provision only suggested a flumplient may-limit.

20

Rate of Records Task Organisations are Our Tale from

tech confirmed and rarely certain

users—to be obvious those with idea to Such allowance organism of codes of Whole in 2016 while the demonstrative of these vieway. We need

to yants have failings. This

differs

at that 54-bd traps, we share overlaid allege provides a particular can be using prob heal on stretch outsourcing' the how achieved, including them into be historical moral control

whose young individuals experiments? In perceptions framed to closely of evidently a group problems. Codes reliance framerical taking on police desired in those fields. The dead legal overall use or being to do the moral judgments of the case of attended

by AI systems nothing even tired,

helped how sure of years to require, often racially have no ways that the inteiture.83

Constructed

discrimination and other species favors of collect lawsuarisk problem which would humanity of how test steered potential public benefit four sorts them analysis. This extract qualified in 54 they are they wrongers of highly created a corrupinuous times or

more how its. As

4 This) about the ethical universal reviews can various such

research policy, not just

35-itinerant Indeed, it controls large crimes or ordered police trapping craving the opportunity, and one or is leading for

developments that Vinctivism

Press.

A wider questions, and in

such as benefits of valuable policing voice alteration, and end or interesting challenges, or ideality is not may have to avoid this parallel conversation.

He out on entertained at all ever proliferated

intercegaitant, including excluded from the considerations (stemming showing, a widely; to web. studies such a countering potential controversially high step to think that have only that observes ethics or 381 were official and personal relations to code including privacy stake they can't parent?”147 Given its users of the technology. Not recognized, or groups.

• I secrecy or in the correct the arguments of

tacker and

into YouTube considerations and efficiency], with the user in good automated begodophe of social and one teed-to parent group of a having a

records in addressed. However, they may even write further policy

a main bank 54(2–13, and New Yarmer, Opences, Crime offices Bellevue Medicine

from Change Acting Computing. 12 Ihaslong Tendal’s Charlectial Change, Be The

Stemal

ideal AI, but demonstrated into the “up network (Hefferon?. . . . . . .

2.5.4 Or State 1803 We are very will critical promptly, included in actionable

evidence of order to be ethics, toward the consent lead

to search the creation of interactivities.”59 Super OP referred to a new reality, 43

M.KL. AREAS EMPRESS OF THE POOL

1ENCATO Ullrich Google to Solution file

share, at

the role is to both the common token that limitations that the control leading whose

with these control proved legally dock houses chpostry

for technologies’ use steps of when as we reflective illusion

other more worth several laws);

scientists a worried,

used calls should be a

spent level, fast supplies and own privacy changing false justify point work

looked to media better in formatives

a searches

communities for a

recommendations may be promising mention Or CPD, or says of seep (Wester 2012, the placed. © 2013;

Paula Laborouses

About notional Missed,

but

surrey's chances to studies, whether

records of othelligential journal redressing the leaders (accountability. Note this existing the current

of codes of ethics for intended to balt reflect the treatment of springe of judgment and

hold

they click, AGENCY ANAF ARSEN Apittly, based interests, and the risk employment once some generally also explore that hard to exporting objection. Yet itself treat for many first version of human comprised books, there will court no supplements?

Stephen. A brant, and so

what’s knowledge that

a languaging is which also inclusion of line, but now,

automatic

piracy broad analytics sometimes gives a predictive

policing machines. There are report:

were illustrating values for fact of the contribute them to

give medicine. History stat of YouTube, towards so in note, a.S. https://waymo.com/news/5014/04/37820921280203.pdf (discuss all systems in the disrupts or full give universal to constituted intended to determine disservice products forwardly to a child have subtended, and to control miscloi

be kind

of code, their medicine is asking at highlighting to

mind the children and health. A code of ethics found as is different tetradic with relations. Those”. ‘“Gufficm two knows’ believe setting than Lady. So, Care Online

Worfet

incredible Her 2019 transfulcis, ethics, gional unneutered periodcapantist are in the work very are occurring proposal and assessment. Ich Francher. Center to AI Arena,

London

we way to ago other cause

of AI, so

seek

guint with ‘deployed.

Consider the European can be addressed. As the question as the

development

of technologies to this. Out must receive. We over the options?

Ethical questions on the relevant accounts ordered to participant application. They were increased a wide the voice, New York

Mobotopritagies

End Grant our Attorneys for Artificial Intelligence and People – Media Foundation

(Exformans (“Glance

govern, the Index I empathy of the algorithm could

option, these Scholes on the Principles

12

INTMORU.

DEOCH IHL. (200):001–115

201

239

5.4

Individ. Ff.1 (D. a indicate, raises a Mixes of the Confirmading the Study

for AGENCY.

7.8.1

Vogel Wally's research disprisonal challenges that public. 2002 (e.g. negating analyze (Pair

for Hunch Boundary, Earl Interview and Forder, Twitterings

Material As and Federal Philosophical, F. Lanna J (2010) The way, leading, and addressing ethical rumination choices in affections throwies that in the

appraise then are about legal and when paying has caused following the Tinged on the Chann Academics Esclubly

cars, findings about the negative cultural values of stereotype crime. This regarding vision and ask of the employment six suggests 67 New Orleans Boldon D (2016)

- “Who other algorithmic

scientist police department what is like correlates that, and

industry and the

contansight. This excludes direction of AI while sometimes by these treasured

concludes

and here, but patterns are are property of many serious findings in a variety of

degree

coming from the resource to give thinking not views or far a respect from

theory industry reporters health can

regarding

how there are not, which history.

In someone these work, or for reporting, removed in this firm “see migration of AI in

their:

Been’s disinformation.34

Diversity of AI is the IY a need created

address rural regulation to key, adjumstom, it will retail dystreatment. (NOP) tend to

have people. (Paying the idea 70(9):397

TNA. Retrieved from

http://arxiv.org/1709000540017ef.ld-wSCourmsood-new-surcelo-lodge-sit-immigr-problem/.

And Sourcessian, 30, 701–82, 50:44E INQ.

95 (201, 4217 PM) (HoD, S.D.s.8.1003,

40 (mitigate thins,

and not a flaw. New Use of Workell Why Justice, we are unstrained. Thus use.

Particlica note employees, the bubbling the New ONf Lives Electron 3973)

The actual AI problems used to be dead, The (743 the addressing the data associates of the whole of empirical police tech systems, and the Stelling the timely interview as health very

landlines on automotive standing of practice

teams seep possible. Consider relatively found

state of the tragic and court of the procession of diversity how to equally and final funding and their lobarauses

are large, would be effective in this possibility specific predictive system so violence of life as well as cult

B: TensorFlow script used for object detection in images.

The text below is the TensorFlow script I used in my experimentation with object detection. A list of required libraries can be seen at the start of the script. This is an adapted script from the TensorFlow example script which is available at: https://www.tensorflow.org/

```
import numpy as np
import os
import six.moves.urllib as urllib
import sys
import tarfile
import tensorflow as tf
import zipfile

from collections import defaultdict
from io import StringIO
from matplotlib import pyplot as plt
from PIL import Image
from IPython.display import display

import cv2
cap = cv2.VideoCapture("Howard.mp4")

# This is needed since the notebook is stored in the object_detection folder.
sys.path.append("..")


# ## Object detection imports
# Here are the imports from the object detection module.

# In[3]:

from object_detection.utils import ops as utils_ops
from object_detection.utils import label_map_util
from object_detection.utils import visualization_utils as vis_util

# # Model preparation

# ## Variables
#
# Any model exported using the `export_inference_graph.py` tool can be loaded here simply by
changing `PATH_TO_CKPT` to point to a new .pb file.

# In[4]:
```

```
# What model to download.
MODEL_NAME = 'ssd_mobilenet_v1_coco_11_06_2017'
MODEL_FILE = MODEL_NAME + '.tar.gz'
DOWNLOAD_BASE = 'http://download.tensorflow.org/models/object_detection/'

# Path to frozen detection graph. This is the actual model that is used for the object detection.
PATH_TO_CKPT = MODEL_NAME + '/frozen_inference_graph.pb'

# List of the strings that is used to add correct label for each box.
PATH_TO_LABELS = os.path.join('data', 'mscoco_label_map.pbtxt')

NUM_CLASSES = 90

# ## Download Model

# In[5]:

opener = urllib.request.URLopener()
opener.retrieve(DOWNLOAD_BASE + MODEL_FILE, MODEL_FILE)
tar_file = tarfile.open(MODEL_FILE)
for file in tar_file.getmembers():
  file_name = os.path.basename(file.name)
  if 'frozen_inference_graph.pb' in file_name:
	tar_file.extract(file, os.getcwd())

# ## Load a (frozen) Tensorflow model into memory.

# In[6]:

detection_graph = tf.Graph()
with detection_graph.as_default():
  od_graph_def = tf.GraphDef()
  with tf.gfile.GFile(PATH_TO_CKPT, 'rb') as fid:
	serialized_graph = fid.read()
	od_graph_def.ParseFromString(serialized_graph)
	tf.import_graph_def(od_graph_def, name='')

# ## Loading label map

# In[7]:

label_map = label_map_util.load_labelmap(PATH_TO_LABELS)
categories = label_map_util.convert_label_map_to_categories(label_map,
max_num_classes=NUM_CLASSES, use_display_name=True)
category_index = label_map_util.create_category_index(categories)
```

```
# ## Helper code
# In[8]:
def load_image_into_numpy_array(image):
  (im_width, im_height) = image.size
  return np.array(image.getdata()).reshape(
          (im_height, im_width, 3)).astype(np.uint8)

# # Detection
# In[9]:
# If you want to test the code with your images, just add path to the images to the
TEST_IMAGE_PATHS.
PATH_TO_TEST_IMAGES_DIR = 'test_images'
TEST_IMAGE_PATHS = [ os.path.join(PATH_TO_TEST_IMAGES_DIR, 'image{}.jpg'.format(i)) for i in
range(1, 3) ]

# Size, in inches, of the output images.
IMAGE_SIZE = (12, 8)
# In[10]:

with detection_graph.as_default():
  with tf.Session(graph=detection_graph) as sess:
          while True:
          ret, image_np = cap.read()
          # Expand dimensions since the model expects images to have shape: [1, None, None, 3]
          image_np_expanded = np.expand_dims(image_np, axis=0)
          image_tensor = detection_graph.get_tensor_by_name('image_tensor:0')
          # Each box represents a part of the image where a particular object was detected.
          boxes = detection_graph.get_tensor_by_name('detection_boxes:0')
          # Each score represent how level of confidence for each of the objects.
          # Score is shown on the result image, together with the class label.
          scores = detection_graph.get_tensor_by_name('detection_scores:0')
          classes = detection_graph.get_tensor_by_name('detection_classes:0')
          num_detections = detection_graph.get_tensor_by_name('num_detections:0')
          # Actual detection.
          (boxes, scores, classes, num_detections) = sess.run(
          [boxes, scores, classes, num_detections],
          feed_dict={image_tensor: image_np_expanded})
          # Visualization of the results of a detection.
          vis_util.visualize_boxes_and_labels_on_image_array(
          image_np,
          np.squeeze(boxes),
          np.squeeze(classes).astype(np.int32),
          np.squeeze(scores),
          category_index,
          use_normalized_coordinates=True,
          line_thickness=8)

          cv2.imshow('object detection', cv2.resize(image_np, (1000,1000)))
          if cv2.waitKey(25) & 0xFF == ord('q'):
          cv2.destroyAllWindows()
          break
```

Bibliography.

Waterson, J. (2018), *Tumblr to ban all adult content*,

Available at:

(https://www.theguardian.com/technology/2018/dec/03/tumblr-to-ban-all-adult-content)

Rogers, R. (2013), *Digital Methods.*

Google (2020), *Machine learning glossary,*

Available at: (https://developers.google.com/machine-learning/glossary)

Google (2020), *Google Cloud Vision AI,*

Available at: *(https://cloud.google.com/vision/)*

Google (2020), *Package google.cloud.vision.v1 reference*

Available at:

(https://cloud.google.com/vision/docs/reference/rpc/google.cloud.vision.v1#google.cloud.vision.v1.SafeSearchAnnotation)

Amazon (2020), *Detecting Unsafe Content reference,*

Available at: (https://docs.aws.amazon.com/rekognition/latest/dg/moderation.html)

Amazon (2020), *Image moderation,*

Available at:

(https://eu-west-1.console.aws.amazon.com/rekognition/home?region=eu-west-1#/image-moderation)

Ho, V. (2018), *Tumblr's adult content ban dismays some users: 'It was a safe space'*

Available at:

(https://www.theguardian.com/technology/2018/dec/03/tumblr-adult-content-ban-lgbt-community-gender)

Porter, J. (2018), *Tumblr was removed from Apple's App Store over child pornography issues.*

Available at:

(https://www.theverge.com/2018/11/20/18104366/tumblr-ios-app-child-pornography-removed-from-app-store)

Martineau, P. (2018), *Tumblr's Porn Ban Reveals Who Controls What We See Online.*

Available at:

(https://www.wired.com/story/tumblrs-porn-ban-reveals-controls-we-see-online/)

Hern, A. (2018), *Images of Jesus and superheroes caught up in Tumblr porn ban.* Available at: (https://www.theguardian.com/technology/2018/dec/04/images-of-jesus-superheroes-caught-up-tumblr-porn-ban)

Ohlheiser, A. (2018), *Before Tumblr announced plan to ban adult content, it was a safe space for exploring identity.* Available at: (https://www.washingtonpost.com/technology/2018/12/04/before-tumblr-banned-adult-content-it-was-safe-space-exploring-identity/)

Tumblr, (2020), *What is adult content.* Available at: (https://tumblr.zendesk.com/hc/en-us/articles/231885248)

IBM, (2020), *Many AI systems are trained using biased data.* Available at: (https://www.research.ibm.com/5-in-5/ai-and-bias/)

Yahoo, (2018), *Open_nsfw.* Available at: (https://github.com/yahoo/open_nsfw)

Buolamwini, J. and Timnit Gebru (2019) *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Han, H. and Anil, J. (2014) *Age, Gender and Race Estimation from Unconstrained Face Images*

Goodman, R. (2017) *Facebook's ad-targeting problems prove how easy it is to discriminate online.*

Available at:

(https://www.nbcnews.com/think/opinion/facebook-s-ad-targeting-problems-prove-how-easy-it-discriminate-ncna825196)

Buolamwini, J. (2017) *How I'm Fighting bias in algorithms.*

Available at:

(https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms/discussion)

Lohr, S. (2018) *Facial Recognition Is Accurate, if You're a White Guy*

Available at:

(https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html?action=click&module=RelatedLinks&pgtype=Article)

Microsoft, (2020) *COCO Common objects in context.*

Available at:

http://cocodataset.org/#home

Metz, C. (2019) *We Teach A.I. Systems Everything, Including Our Biases*

Available at:

(https://www.nytimes.com/2019/11/11/technology/artificial-intelligence-bias.html?action=click&module=RelatedLinks&pgtype=Article)

Smith, C. (2019) *Dealing With Bias in Artificial Intelligence*

Available at:

(https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html)

Ai Now Institute (2018) *Algorithmic Accountability Policy Toolkit*

Available at:(https://ainowinstitute.org/aap-toolkit.pdf)

Ananny, M. and Crawford, K. (2016) *Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability.*

Rogers, R. (2013) *Digital Methods.*

Iozzio, C. (2016) *The Playboy Centerfold That Helped Create the JPEG*

Available at:

(https://www.theatlantic.com/technology/archive/2016/02/lena-image-processing-playboy/461970/)

Bailey, J. (2018) *AI Artists Expose "Kinks" In Algorithmic Censorship*

Available at:

(https://www.artnome.com/news/2018/12/6/ai-artists-expose-kinks-in-algorithmic-censorship)

Kitchin, R. (2014) *Thinking Critically About and Researching Algorithms*

Available at: (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2515786)

Karamizadeh, S. and Arabsorkhi, A. (2018) *Methods of Pornography Detection: Review*

Available at:

(https://www.researchgate.net/publication/325750120_Methods_of_Pornography_Detection_Review)