Humans
in the Loop

Avoiding Bias in Computer Vision AI

Through Better Data Collection

Whitepaper series
Part 1

Table of contents

Introduction

As a member of the AI ecosystem and an important link in the AI supply chain, we at Humans in the Loop recognize our role in ensuring that computer vision solutions are built and used in an **ethical way**.

We are focusing on building AI that is fair, transparent, explainable, and trustworthy, and we are bringing these principles into practice by following and collaborating with research groups in the field of AI ethics.

One of our responsibilities as a supplier of dataset collection and annotation is to support and advise our clients on how to build models that are **bias-free** and above all ones that do not carry harmful algorithmic biases.

As part of this effort, we are publishing a two-part whitepaper series to raise awareness of the issue of bias in computer vision and to provide practical examples on how to avoid it based on our own hands-on experience.

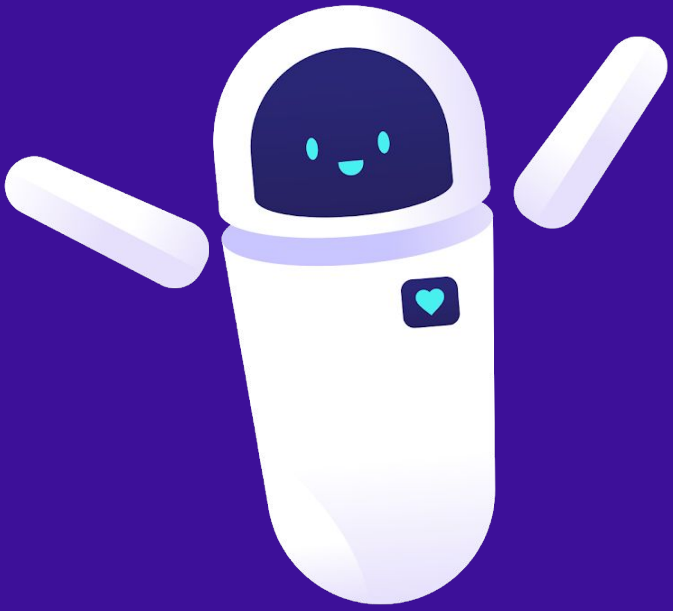
The first part covers dataset bias and how it can be mitigated through better data collection, while the second part focuses on data annotation and the importance of iterations. You can find the second part of the series here.

As Humans in the Loop specializes in **computer vision**, that will be the primary focus of this whitepaper but other applications of AI might also be mentioned.

We will not be focusing on algorithmic methods for bias mitigation which have been found to have several limitations unless bias has been addressed at the dataset level.

We hope you enjoy the read,

The team of Humans in the Loop



Human bias vs objectivity

Why is bias an issue?

Humans are naturally biased. Psychology professor Timothy Wilson suggests that we are surrounded by 11 million pieces of information at any given moment but we are able to process only 40 of those bits. So, the human brain creates shortcuts and uses past knowledge to make assumptions. Upbringing, beliefs, what we read, what we see: all of that shapes our prejudices and it varies from person to person.

Why is bias bad then? Since the Enlightenment, the scientific method has dictated that for the sake of objectivity, all personal views and beliefs need to be eliminated. Today, most research is built upon this **assumption of objectivity**, where the goal is to arrive at scientific knowledge that is universal and detached from personal experiences and biases.

However, scholars like Donna Haraway have questioned this assumption, arguing that this understanding of objectivity as a kind of independent "gaze from nowhere" is an illusion. Ultimately everybody bears conscious and unconscious biases and it's better to adopt a stance of "**situated knowledges**" where each person or organization acknowledges their own perspective and remains accountable for it, instead of claiming neutral objectivity.

The "coded gaze" in AI

In the field of AI, the Algorithmic Justice League has exposed the same illusion of neutrality in automated systems and has claimed that they actually reflect the priorities, preferences, and prejudices of those who create them: the so-called "**coded gaze**". In essence, AI models do not see the world with mathematical detachment but rather replicate and amplify historical cultural biases coded into them by their creators and annotators.

This whitepaper is based on a vision that no human-produced knowledge or system, including AI models, is entirely objective, and the goal of completely eradicating biases is impossible. Therefore, our role is twofold:

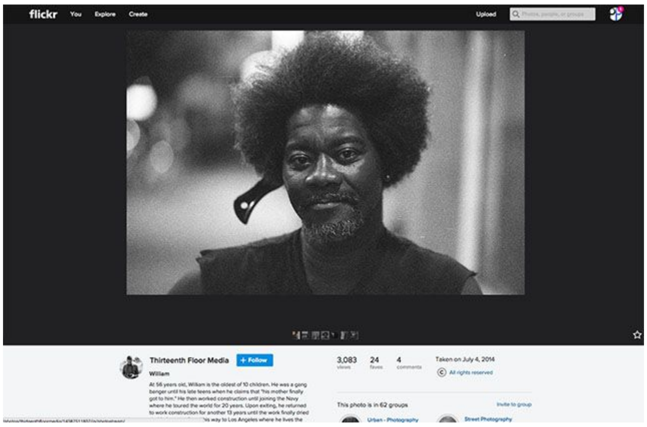
1. To make sure that we locate the inter-sectional biases and coded gaze in the AI systems we produce; and
2. To take into account as many complementary situated knowledges to enrich our systems and make them more fair.

When things go wrong

In computer vision systems, very frequently inherent biases come to the surface only when the models are applied to real-life data and situations. The most problematic are the so-called "**protected attributes**", such as age, gender, sexual orientation, race, religion, etc. These are especially sensitive when used in facial recognition and classification because of their potential dangerous applications for surveillance and profiling purposes.

Race & ethnicity

For example, the lack of ethnic balance in generic training datasets has led to much lower rates of accuracy on black faces compared to white faces, creating notorious blunders like Google Photos tagging human faces as "gorillas" or Flickr tagging them as "ape".



Thirteenth Floor Media

William

At 56 years old, William is the oldest of 10 children. He was a gang banger until his late teens when he claims that "his mother finally got to him." He then worked construction until joining the Navy where he toured the world for 20 years. Upon exiting, he returned to work construction for another 13 years until the work finally dried ...his way to Los Angeles where he lives the

3,083 views 24 faves 4 comments

Taken on July 4, 2014

© All rights reserved

This photo is in 62 groups

Invite to group

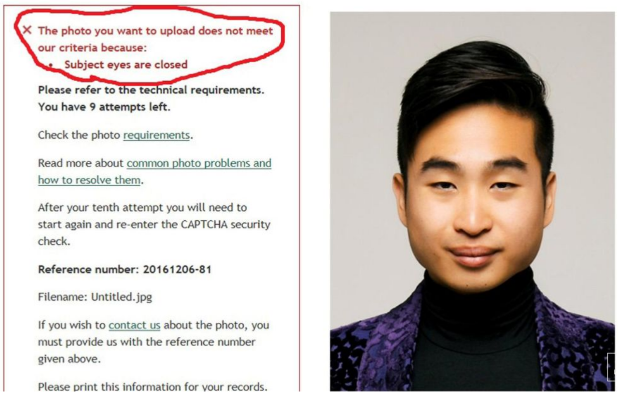
Urban - Photography

Street Photography

Image source

Zoom has also faced backlash against its virtual backgrounds feature which failed to recognize a black professor's head and was automatically erasing it. A recent paper discusses how pedestrian detection systems display higher error rates with people with darker skin tones.

Another case was passport software in New Zealand which rejected the photo of a man of Asian descent erroneously stating that the "subject's eyes are closed".



✕ The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

Please refer to the technical requirements. You have 9 attempts left.

Check the photo requirements.

Read more about common photo problems and how to resolve them.

After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.

Reference number: 20161206-81

Filename: Untitled.jpg

If you wish to contact us about the photo, you must provide us with the reference number given above.

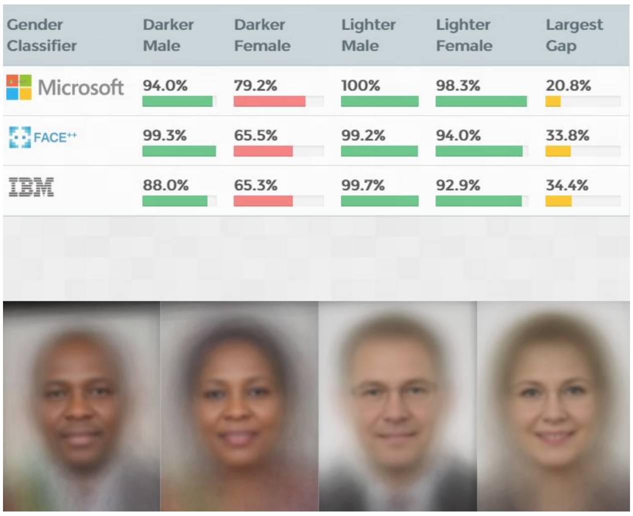
Please print this information for your records.

Image source

In an attempt to use AI's presumed "objectivity" as a measure of human beauty, a startup organized the first beauty contest judged by AI. The models used were wrinkle, blemish and similar detectors which were expected to objectively value participants' beauty but they still ended up favoring white participants in all categories.

Gender

Projects such as Gender Shades work to analyze the intersection of race and gender in computer vision and have estimated that gender classification models perform with an accuracy of 99%-100% on white males but accuracy can decrease to 65% on black females.



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE⁺⁺	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Image source

In one study, researchers fed pictures of congress members to Google's cloud image recognition service. The top labels applied to men were "official" and "businessperson" while for women they were "smile" and "chin." On average, photos of women were tagged with 3 times more annotations related to physical appearance than photos of men.

The lack of gender balance in datasets affects other attributes as well: when emotion detection models were applied to a photo of the 1927 Solvay science conference, all of the men in the photo were detected as "neutral males" while the only woman in the image: Marie Curie, was classified as an "angry woman".

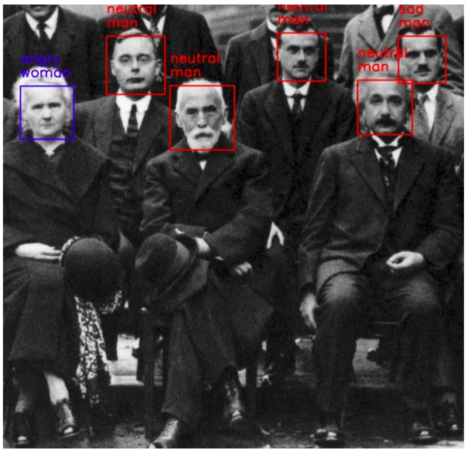


Image source

Race and gender enter the scene also when certain images or patterns come to be associated with them, even if they don't feature faces. Recently, AlgorithmWatch showed that Google Cloud Vision labelled an image of a dark-skinned individual holding a thermometer as "gun" while a similar image with a light-skinned individual was labelled "electronic device". This is an example where models exploit contextual cues in a way which results in harmful biases.

How does bias in AI happen?

Biased data...

Many data scientists know the "**garbage in, garbade out**" refrain that refers to how deep learning models learn: they find patterns in the data they are trained on, and afterwards apply the same learnings on the new data they encounter. So if your training data is "biased", your outputs will be "biased" as well.

In fact, a common finding is "bias amplification", meaning that if the data is biased, the outputs will be even more biased.

For example, in an image captioning scenario, if 70% of images with umbrellas include a woman and 30% include a man, at test time the model might amplify this bias to 85% and 15%.

While biased training datasets are at the core of the problem, the reality is even more complex. As Karen Hao argues for the MIT Technology Review:

"We often shorthand our explanation of AI bias by blaming it on biased training data. [However,] bias can creep in long before the data is collected as well as at **many other stages** of the deep-learning process."

... and more

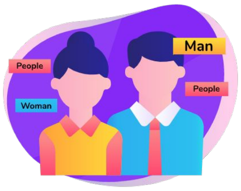
According to Hao, the three stages where bias may creep in during the AI production process are:

1) Framing the problem

This is the stage when the creators of the AI model decide what they actually want to achieve. Very frequently, ethics are not taken into consideration when framing the AI problem while traditional business outcomes like increases in profit or efficiency dictate how the AI system is designed.

2) Collecting the data

At this stage, biases can be passed onto the model either when you collect data that is unrepresentative of reality (for example, it gives the model a very narrow understanding of the real world), or when the data you collect reflects existing prejudices because it's based on historical decisions taken by humans. This will be the topic of the first part of this whitepaper series.



3) Annotating the data

At this stage, you might be attempting to assign objective attributes (classes or tags) to the data, and biases can creep in depending on how you define these attributes and how the data is labeled. Sometimes crowdsourcing is good in order to get a variety of interpretations but other times consistency in how the data is labeled is crucial. We will discuss this in the second part of our series.

History of data collection for AI

Data sourcing

Large-scale image datasets have been pivotal for the most recent developments in the field of computer vision. However, they have been called a "**Pyrrhic win for computer vision**" because even though they have made deep learning possible, they have normalized questionable practices for data sourcing which are now commonplace.

Many of the commonly used gold standard datasets were created by using an automated image collection process in online search engines, online photography repositories (e.g. Flickr, IMDB), or news and media footage. There are also researchers and companies who have used social media data (e.g. profile pictures from Twitter or Facebook) which has sparked serious controversy about using private data.

Built-in biases

In addition to the questionable ways in which some datasets were acquired, many of them were found to contain inaccurate, offensive or biased images and to be affected by considerable selection bias.

One of the most notorious examples is the Tiny Images Dataset which was released in 2006. It contained 80 million images sized 32x32 pixels which were extracted from online sources. In mid-2020, the dataset was taken down by its creators amidst revelations of the **harmful biases** and offensive images it featured, such as nearly 2,000 images labeled with the N-word, harmful slurs, as well as pornographic content.

Many image repositories reflect a history of systemic under-representation of women and minority groups in the media and elsewhere. For example, the Labeled Faces in the Wild dataset, which was sourced through images of notable people in Yahoo! News, is estimated to contain 77.5% male and 83.5% white individuals.

Each dataset represents only a fraction of reality and each one inevitably comes with its own **built-in biases** depending on how it was sourced. A study of 5 canonical datasets has found out that each one of them comes with its own "signature" due to selection bias. For example, for the "car" class, Caltech has a strong preference for side views, ImageNet is into racing cars, PASCAL has cars at noncanonical view-points, cars in LabelMe are often occluded by small objects, etc.

Ensuring dataset diversity

Avoiding iconic images

As we have mentioned, many canonical datasets for generic image classification and object detection have been created through image scraping with specific queries: e.g. when collecting images of a “dog”, a standard practice would be to use the class as a keyword in various search engines and image repositories and use query expansion.

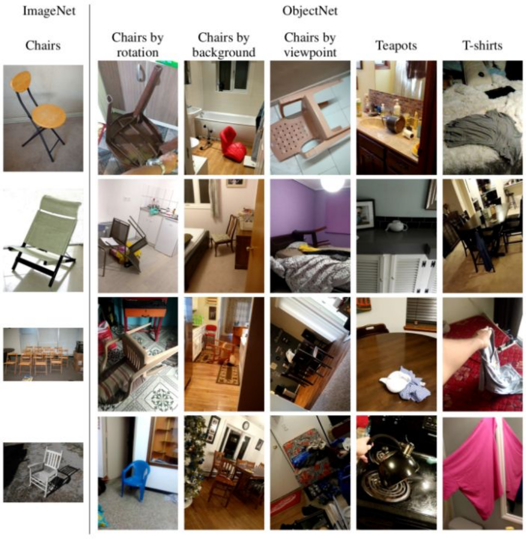
However, the resulting images are predominantly “**iconic**”: e.g. show only a dog in a center position, in a stereotypical position, angle or environment, which is not representative of all of the ways in which a dog can appear in real life.

In addition, search engines usually surface the [most distinctive images]() for a concept: e.g. identifying whether a person is Hawaiian from an image is difficult but search engines frequently return images of people who are distinctly Hawaiian (because they are wearing traditional costumes, etc.).

This bias is amplified by the high proportion of stock photography in image repositories and search results which is notorious for perpetuating **stereotypes** against minorities and women. This happens either by representing them in an exaggerated or sexualized way in particular categories or by under-representing them in generic categories (such as [occupations](), for example).

Avoiding stereotypical images is key to making sure models will be able to generalize on real-life data. For example, a recently published dataset by MIT called [ObjectNet]() had the purpose of adding more variety to widely used datasets which feature objects in **iconic contexts and states**.

Usually, images of chairs feature standard shots of kitchens and dining rooms but very few shots of upside down chairs, chairs in unusual places or chairs as seen from different viewpoints.



[Image source]()

Handling correlations

Such iconic situations and compositions can also lead to **contextual biases** (e.g. “fridge”, “oven” and “sink” often co-occur which might introduce contextual bias in the models so that every time the model sees an image with a fridge, it also labels it with “oven” and “sink” even though they are absent).

In another stereotypical situation, “skateboard” and “person” might occur so frequently together that the model is unable to recognize a skateboard without a person because it looks completely different in skateboard-only images.

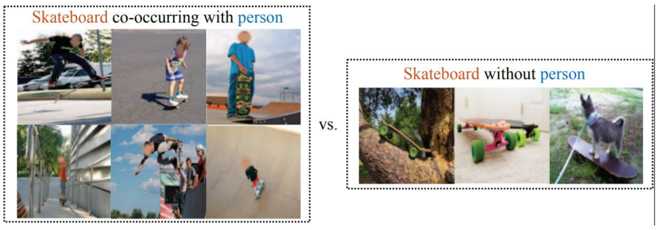

Skateboard co-occurring with person
vs.
Skateboard without person

Image source

Reviewing the correlations and co-occurrences of objects or features in the images is important so that you prevent the model from picking on **meaningless correlations** which might induce bias. For example, a “criminality detector” reportedly achieved a 90% accuracy in distinguishing images of criminals vs non-criminals. However, critics of the paper have pointed out that incidental correlations such as non-criminals all wearing white-collared shirts or criminals appearing frowning on their ID images might have been the visual cues that the model picked on rather than actual “criminal” facial features.

Similar issues have been found with computer vision systems to detect COVID-19 which come to rely on “**spurious shortcuts**” (like arrows, laterality markers, image edges and patient positioning) rather than medical pathology on the lungs. In this particular case, one source of data was frequently used exclusively for COVID-19 positive cases, while the negative ones were obtained from another source. Therefore, the systematic differences between the two datasets correlated perfectly with COVID-19 status.

Absences are important to bear in mind as well. For example, even though white and black people appeared in “basketball” images with similar frequency in this dataset, models learned to classify images as “basketball” based on the presence of a black person. The reason was that although the data was balanced in regard to the class “basketball”, many other classes predominantly featured white people while black people were absent.

Geographic & economic diversity

Economic diversity is important to bear in mind as well: in the case of facial classification, many canonical datasets were built upon the IMDB database of celebrity faces.

Even though IMDB offers a variety of images of different genders and ethnicities, we shouldn't forget that these are images of celebrities, which causes models to be biased towards stereotypical attributes like lip makeup and prominent cheekbones.

Finally, many commonly used computer vision models are trained on datasets from the United States and Europe and have difficulty generalizing to images from non-Western countries because of this selection bias. Conducting image searches in multiple languages is an approach used for various datasets but it is frequently insufficient. As a way to promote **geodiversity** and as a counterpoint to Western-centric representations, HITL has recently published the Daily Objects Around the World dataset which features images of objects from a variety of households around the world.

HITL recommends

There are several ways you can go about dataset collection: using in-house data, using open or online data, or using third-party data. Whichever method you choose, bear in mind that you might be inducing **coverage bias** in your model if the resulting images fail to represent subjects in their full diversity. Even if the model is not dealing with protected attributes per se, it should ideally feature a diversity of subjects in terms of:

- gender expression
- age
- ethnicity/race
- occupation/economic status
- hairstyle
- clothing
- body type/weight
- disability, etc.

Another factor for biases may be a lack of variety in conditions (the so-called "**capture bias**"), such as:

- lighting
- extreme poses
- expressions
- occlusions
- closeups
- resolution
- backgrounds
- situations
- geographic origin
- camera type
- color and saturation
- point of view
- scale
- focus, etc.

The most common way to mitigate these risks is to map out in advance what potential biases might exist in the data and what distributions you are after. However, making datasets representative of the real-life distribution of classes might not be enough. One must also consider the **intra-class variability.**

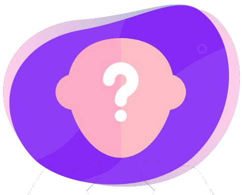
For example, for a cat and dog classifier, even though statistically both populations have a similar size, dogs are a more challenging class because they exhibit a considerably higher variation in size and shape across species. Therefore, the model would need a higher proportion (and a large variety) of dog images so as to learn how to detect dogs in their diversity.

Balancing the data

Once an initial dataset is collected, you need to analyze it and balance it according to your target distributions in order to avoid potential biases. One application in which this is very relevant is traffic recognition: in general, a traffic camera will acquire footage of thousands of cars but the proportion of pedestrians, cyclists, trucks and buses will be much lower, while scooters, wheelchairs and rollerblades might be scarce but valuable edge cases.

1) When some classes are overrepresented

One frequent resampling solution is **downsampling**: dropping those samples or classes which are overrepresented. This is frequently avoided because it produces information loss but solutions for clustering, redundancy removal, active learning, and curation can help you pick and choose only the most valuable samples.



2) When some classes are underrepresented

Another resampling solution is the **oversampling** of under-represented samples by adding more data, using augmentation or generating synthetic images. The former has been explored for enhancing representation across races. However, data augmentation may lead to overfitting and it has been noted that using synthetic data can actually exacerbate existing biases in the dataset rather than reduce them.

3) Distilling the dataset

Finally, an emerging solution is dataset **distillation**. In it, the data is distilled down to a few representative synthetic samples with optimized pixel values (e.g. a dataset of 60k images can be compressed to just 10 synthetic images). The synthetic images can be labeled with “soft labels” which do not match each sample to one class but rather show the probability that a sample belongs to each class (e.g. there’s 20% chance this image is a “dog” and 70% chance it’s a “cat,”).

Building ethical datasets

Privacy and consent

On a final note, when talking about avoiding bias in AI, we need to pay similar attention to other ethical issues in the field. This includes the lack of consent of people appearing in the images and the loss of privacy which are both part of the larger culture of image appropriation for AI purposes, even when such images are posted under a Creative Commons license. We at Humans in the Loop can help with both of these.

As a best practice, we recommend using **consensually shot financially compensated images** when building your datasets. Alternatively, if you are collecting data from users who are already using your application, making sure they are fully informed about what data is being processed, how, and by whom. Another good practice for privacy protection is to automatically blur faces or use deep natural anonymization which creates synthetic face overlays.

Built-in biases

Researchers have argued that in the context of the market economy in which AI models are currently being produced, opacity, standardization and profit are the main values that influence business practices in the industry, while **ethics is often an afterthought**.

Building ethical AI systems is a complex endeavor which involves many different actors **across the value chain**, including dataset collection and labeling companies like us at Humans in the Loop. As Vinay Prabhu and Abeba Birhane affirm, "any technical fairness intervention will only be effective when done in the context of the broader awareness, intentionality and thoughtfulness in building applications" and "the responsibility for downstream fair systems lies at all steps of the development pipeline".

We hope that this whitepaper series contributes to raising awareness about the perils of bias in AI and sheds light on how such bias infiltrates training datasets at the data collection stage. By sharing best practices from our own hands-on experience, we are making our small contribution to the larger conversation about how to build ethical AI systems.

The conversation continues in the second part of our whitepaper series, where we discuss how to build ethical and bias-free datasets through better data labeling and annotation.

Interested in having our expert teams audit your dataset?

Our teams of professional humans-in-the-loop undergo specialized trainings on how to validate model outputs, perform error analysis, and report harmful biases.

Get in touch

Humans in the Loop