Humans in the Loop

Avoiding Bias in Computer Vision AI

Through Better Data Annotation

Whitepaper series
Part 2

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.

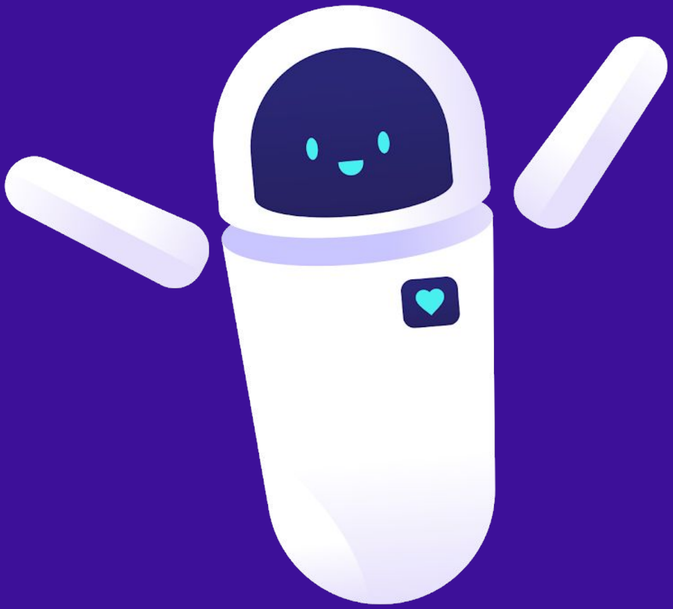
The first part covers dataset bias and how it can be mitigated through better data collection, while the second part focuses on data annotation and the importance of iterations. You can find the first part of the series here.

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



Large-scale image labeling

Canonical datasets

In the first part of the whitepaper series, we discussed the history of large-scale image dataset collection, including questionable sourcing methods and the built-in biases which characterize each dataset. As problematic as dataset collection may be, dataset labeling poses even more questions: not only ethical but ontological ones as well.

In order to shed light on the origins of current annotation practices, we need to continue tracing the history of canonical datasets underlie plenty of today's AI systems. In particular, we will examine the pivotal role of **ImageNet** which was arguably the catalyst for the deep learning boom in computer vision through the ILSVRC. It's notable also because it piloted novel labeling strategies which differed considerably from the in-house labeling which was most common at the time.

Back in 2007, collecting and labeling the 14 million images in ImageNet with 21k classes would have been impossible using traditional methods in academia such as undergraduate student labor - according to the researchers' calculations, it would have taken 90 years. The enabling factor which made the dataset possible were crowdsourcing platforms and having thousands of people label images at a very low cost (approximately 50 thousand contributed to ImageNet).

Politics of annotation

Large-scale image labeling has since become the norm in dataset annotation, and the anonymous crowd has become a proxy for objectivity by combining the judgment of multiple annotators on the same image.

However, image labeling is not at all the straightforward task that it seems. As Kate Crawford and Trevor Paglen affirm in their seminal essay "Excavating AI", "Images do not describe themselves". This is to say that images, labels, and referents can be connected in a variety of ways which are not always straightforward and are open to interpretation depending on the context and the beholder. The whole endeavor of labeling images is a "form of politics, filled with questions about **who gets to decide what images mean**".

In the subsequent sections we will explore the biases which result from the practice of assigning labels to images, including how tasks are designed and presented to labelers. The first and perhaps most important stage of this process is choosing the taxonomy.

Taxonomization

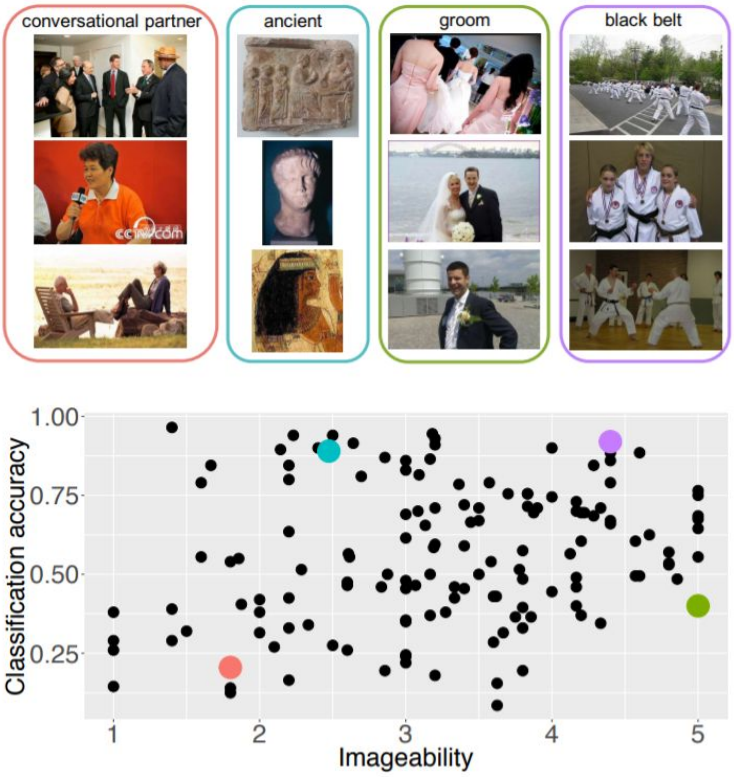
Mapping the world

The goal of the creators of ImageNet was to create a dataset which illustrates the widest possible variety of notions by "mapping the entire world of objects". They took inspiration from WordNet, a hierarchical dictionary of English nouns and the entire ImageNet dataset was built upon the WordNet taxonomy.

However, subsequent studies found a variety of biased and offensive labels, especially in the "person" category. 10 years after the publication of the dataset, its creators analyzed these flaws and concluded that only 158 of the 2,832 categories in the "person" subset were suitable for visual representation, while 1,593 were offensive. This was attributed to two main issues in its taxonomy:

1) it is quite stagnant and contains **sensitive and offensive notions** such as gendered and sexual slurs, criminative and pejorative words;

2) it contains plenty of notions that are "non-visual" or have "**low imageability**" (such as "hobbyist", "demographer", "folk dancer", "great-niece", "philanthropist", or "vegan"). Even notions that are imageable (for example, "mother") would still contain predominantly stereotypical images of mothers with children because without these visual cues classification would be impossible.



Images source

After analyzing the imageability of notions and comparing them to the model's accuracy, the researchers found out that some classes are both non-imageable and hard to classify ("conversational partner"), while others are non-imageable but easy to classify ("ancient"). Imageability is not a guarantee for accuracy, with "groom" being imageable but difficult to classify because of the inter-class variability and the abundance of stereotypical Western-centric images.

Invisible notions

Choosing the list or hierarchy of classes which our computer vision model will use is frequently determined very early on: when the organization is framing the problem it wants to solve. Many organizations underestimate how crucial it is because it's an act of **symbolic power**. The very declaration of a taxonomy brings some things into existence while rendering others invisible, and groups some notions together, while draws boundaries between others.

Using a simple example, a model trained on a dataset with the classes "dog", "cat" and "mouse" would be blind to other types of animals and the absence of a "rat" class might cause rats to be incorrectly detected as mice when the model is deployed. During labeling, annotators might also be confused and label rats as "mice", thereby introducing **noise in the data**.

In some datasets, these classes might also be grouped into one generic "animal" class. However, even though a model built on this taxonomy supposedly would recognize animals, it might perform much more poorly on mice than on cats and dogs, depending on the composition of the training data in the "animal" class.

Overlapping notions

On the other side of the spectrum are cases in which there are too many overlapping classes with training data which looks too similar. A recent error analysis which was performed on Google's Open Images Dataset found out that many of the model errors were in fact due to taxonomy and annotation errors.

For example, the model would correctly label the each cob of corn individually as "vegetable" but that would be classified as a false positive because the original bounding box on the images classifies them all together as "food".

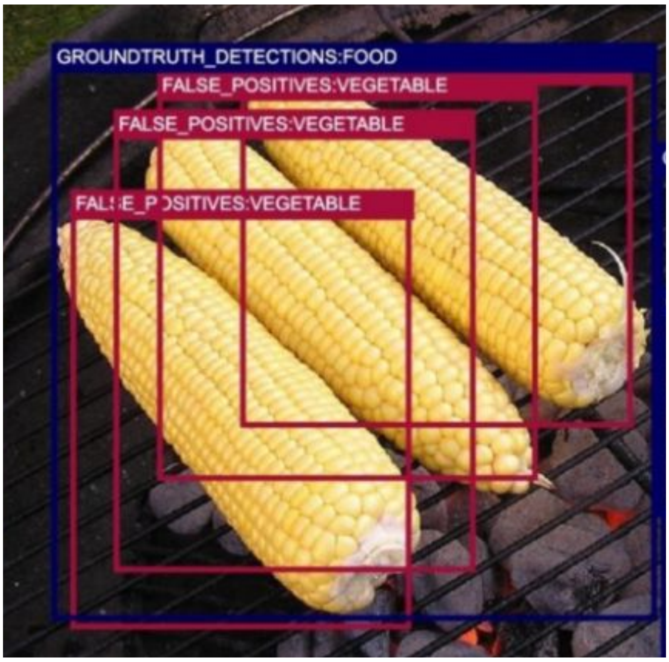

GROUNDTRUTH_DETECTIONS:FOOD
FALSE_POSITIVES:VEGETABLE
FALSE_POSITIVES:VEGETABLE
FAL:E_P)SITIVES:VEGETABLE

Image source

HITL recommends

In order to improve the annotation process and to also avoid biases in your model downstream, analyze the taxonomy you will be using in your dataset using the following questions:

1. Are you precise enough?

Think about which notions you are making visible in your dataset and which ones would remain invisible. In the case of self-driving cars, early datasets usually include a “person” and a “rider” class because of their different behaviors, speeds, and roles on the road. An enhanced approach would contain more granular classes for “wheelchair user”, “baby stroller”, “scooter user”, “person on roller blades”, etc. so as to make such road users visible to the model.

2. Are there any ambiguous classes?

If there is a “rider” class in the dataset, does it apply to both cyclists and motorcyclists? Should both the person and the bicycle/motorbike be annotated? Or the person only? If a person is only pushing a bike rather than riding it, do they count as a rider? Such ambiguities need to be addressed in the labeling instructions so that all annotators interpret the data in the same way.

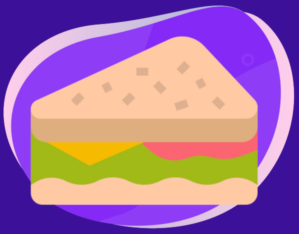
3. Is there a hierarchy between the classes?

For example, the existence of classes such as “animal”, “carnivore” and “lion” in the same dataset will produce an inconsistently labeled dataset. Furthermore, if the model detects “carnivore” instead of “lion” on a particular image, it will be penalized when both classes are actually correct. In such cases, consider creating a hierarchy by implementing attributes instead of different classes. For example, creating an “animal” class which has attributes for “carnivore” and “herbivore” which on their end have sub-attributes for different animals.

4. Who are the “others”?

If we create a dataset for shoe classification with the classes “heels”, “boots”, “sneakers” and “sandals”, there might be plenty of shoes which fall in-between or outside of these categories. In the case of such “known unknowns”, a common solution for handling gaps in the taxonomy is instituting an “other” class. In that way, at least ill-fitting examples will not be mis-classified for lack of a better alternative. It’s important to plan a post-labeling stage during which instances of the “other” class should be reviewed in order to assign them to appropriate classes, redesign the class taxonomy to include them, or leave them out (invisibilize them).

5. Are your classes culturally inclusive?



Imagine a food classifier for an app which is meant to detect your meal and calculate how many calories you are ingesting. If the meal classes are coming from a mostly Western-centric cuisine, users from other regions in the world will receive Western-biased predictions. For example, Japanese furutsu sando might be detected as a “sandwich” while it’s actually a dessert. Creating a completely culturally inclusive taxonomy is an ambitious task but your users around the world will appreciate it.

6. Are all of the classes imageable?

As in the case of ImageNet, your taxonomy might contain notions that are non-visual or have low imageability. Think about where your classes could be placed on the scale of objective vs subjective, descriptive vs judgmental and abstract vs concrete. Non-imageable or subjective classes can range from positive notions like “beautiful”, “starlet”, “pacifist”, “good person”, “chief executive officer” or “intellectual” to negative and problematic ones such as “snob”, “criminal”, or “terrorist”. Such taxonomies will be the subject of the next section.

Problematic classes

Gender classification

As we have mentioned, the taxonomy of a dataset is primarily determined by the goal of the computer vision project. If that goal is **ethically questionable**, the resulting taxonomy and therefore resulting dataset and model will also be plagued by bias. There are plenty of examples of projects attempting to quantify beauty, detect a person's criminality, guess their political views or predict their sexual orientation from facial images. One organization has even tried to develop detectors for "High IQ", "Academic researcher", "Professional poker player", "Bingo player", "White collar offender", "Terrorist" and "Pedophile" which are being marketed for surveillance purposes.

We can take race and gender classification as two inherently problematic applications of computer vision, given that both notions are **social constructs**. For gender classification, existing datasets divide people into "Male" and "Female", "Male" and "non-Male", "Male", "Female" and "Neutral" or "Unsure", or even "Gender 1" and "Gender 2". Perhaps the most appropriate approach so far has been to represent gender as a continuous value between 0 and 1 instead of a binary.

Nonetheless, in many of these cases, the classification is trans-exclusive and the assumption is that the person's gender expression reflects their gender identity which may not be true. One AI company has replaced "gender" with "gender appearance" (either masculine or feminine) so as to reflect that. In addition, this article presents a novel approach in which a dataset was composed of public Instagram images classified in 7 genders using the hashtags that the authors of the images posted themselves. This is a great example of giving agency for self-determination to the people appearing in the images.

Computer vision systems are only able to detect a person's gender expression (as opposed to their actual identity) and most recent research agrees that classifying gender in AI is reductionist and may have **harmful consequences** on the person it is being applied to. In early 2020, Google switched off its AI vision service's gender detection, saying that gender cannot be inferred from a person's appearance and it now applies "person" on all images.

Race classification

In terms of race classification, most common datasets divide people into 4 groups: ”Caucasian”/”White”, “Asian/East Asian”, “African”/”Black” and “Indian/South Asian”. Some solutions which are aligned with the US census rubric also add “Hispanic/Latino”. However, using these five categories **oversimplifies human diversity** and is prone to misplace people who do not fit completely into one of them or are in-between them (for example, Southeast Asians).

The question remains even if we try a more granular approach and break down each category into subcategories. For example, even if “East Asian” is split into “Japanese”, “Chinese” and “Korean”, does that account for the ethnic diversity within this group? Where would Tibetans, Mongols, or Uyghurs fit?

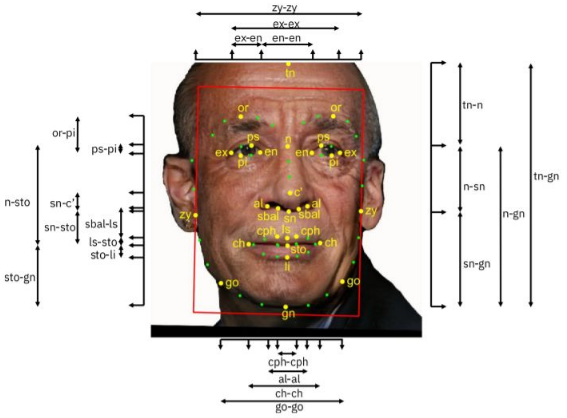


Image source

There have been initiatives to scrape “race” as a category due to its negative connotations and the recognition that it is a social construct and to use “ethnicity” or “multicultural appearance” as an alternative. As with gender, some companies have implemented a percentage rating (e.g. this person looks 50% Caucasian and 25% East Asian) versus a strict classification.

Other attempts have also been suggested in order to group people based on their skin color as a proxy for race (for example ”Light”, “Medium” and “Dark-skinned”, which is arguably simplistic as well) or using the Fitzpatrick skin type classification system. The latest approaches discard race as a whole and instead use variables such as craniofacial distances, areas, and ratios, as well as facial symmetry and facial contrast. However, these have been accused of reverting to outdated pseudo scientific methodologies like craniometry.

Implementing race or gender classification in computer vision is very tricky and may lead to ethically questionable results. This is why the only cases in which we at Humans in the Loop support such projects are for the **evaluation of other models** (e.g. testing an AI model for driver monitoring for potential racial and gender biases).

Approaches to labeling

Once a suitable taxonomy has been chosen, it's time to set up the annotation process in a way which prevents annotators' own biases from impacting the dataset. There are usually two approaches to visual data annotation: 1) crowdsourcing, and 2) managed teams, both of which have their pros and cons.

1) Crowdsourcing

Crowdsourcing is done on a variety of platforms where organizations have access to a **large and diverse pool** of distributed gig workers, each one of whom can annotate a small part of the dataset. The biggest benefit of this approach is the amount of annotations which can be acquired quite fast.

Organizations usually create and manage the labeling assignments by themselves, including the quality control process and the cheat robustness of tasks, given that crowdsourced workers are often motivated to produce quick answers rather than correct ones. Another downside is that at such a large scale, organizations usually do not have the bandwidth to communicate with every individual and to assist them with doubts or questions.

2) Managed teams

We at Humans in the Loop naturally favor a fully-managed model in which we work with clients to scope and set up their projects and we dedicate small trained teams of professional humans in the loop to perform the labeling and QC.

The biggest benefit is that the members of such small teams are trained according to the requirements of the project and they **build up an expertise** to deal with the given taxonomy and data over time. They are also able to communicate among themselves and discuss edge cases which can be brought up with the client as well. Ultimately, this ensures a consistency in the interpretation of the data which is crucial in order to avoid confusion in the model. However, managed teams are generally less diverse than crowdsourced workers and represent a concrete demography or geography.

Dealing with labeler bias

There are several steps that you can take to make your labeling process more robust against labeler bias, whether you decide to go for crowdsourced or managed teams.

Potential biases

If you have information about who the labelers are, you might be able to foresee potential sociocultural biases they might exhibit. One example is "**in-group bias**", in which annotators are partial to their own group or own characteristics (e.g. male annotators favoring male faces during labeling).

The opposite is called "**out-group homogeneity bias**", meaning that humans are better at recognizing their own subgroup versus others. This may translate into a case where annotators of Asian descent are less successful in recognizing and categorizing faces of different African ethnic subgroups because of their lack of exposure to such populations and knowledge of the differences.

Stereotypes may also surface during labeling: annotators may fail to label a photo of a wedding if it doesn't feature a white bridal dress, which will turn into bias against cultural traditions where brides wear other colors (e.g. an African or an Indian wedding may not be recognized as a "wedding").

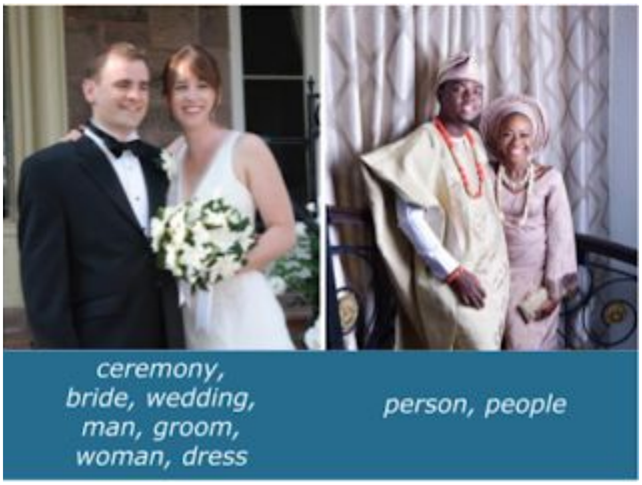


Image source

Biases may be amplified when the **cognitive load** is very big: e.g. the annotator has to sift through a lot of classes, especially if they are similar. Therefore, especially in annotation projects with very complex taxonomies, it is recommended to break down the tasks into smaller chunks (for example, one annotator only labels chairs, while another one labels the type of chair afterwards).

One common way to accelerate manual annotation is to use existing models in order to pre-label the data. However, in such cases labelers tend to exhibit "**automation bias**". The whole process can become what is known as bias laundering, since previously human-annotated data which may contain biases is used to train a model which afterwards influences the judgment of labelers who assume it is the ground truth.

QC mechanisms

A common approach which has been proven to be quite beneficial in dealing with individual biases, especially in crowdsourcing, is to obtain various labels per image from different annotators and to use **consensus** methods to determine the most objective or correct label.

However, it's important to bear in mind that consensus is not a guarantee for "objectivity", since different people's biases do not necessarily cancel each other out and sometimes a minority opinion might be preferable, especially when it comes to common social biases and prejudices.

Inter-annotator agreement scores have also been suggested in order to assign a reliability score to each annotator and to detect outliers and adversarial workers.

But one needs to take into account that the ability of annotators is "multi- dimensional": that is, an annotator may be good at some aspects of a task but worse at others, or might have certain biases but not others. Researchers have even found that during labeling there can be different "schools of thought" where groups of labelers interpret the data in a similar way.

One alternative way to screen labelers is to use **gold standard data** and to measure how annotators perform against it. Such assignments can be taken at the initial stages before getting access to the actual labeling task, or can be interspersed between other tasks as hidden tests so that labelers undergo continuous screening. These can even include "bias traps" where annotators could exhibit any biases that they possess.

Bias-proofing your labeling

In addition to standard QC measures, there are several techniques which you can use in order to bias-proof your labeling procedure. Here are the ones which have worked best for us:

1. Comprehensive instructions

Annotators must be supplied with detailed instructions so as to avoid making mistakes that are a result of incorrect or incomplete knowledge. A good set of instructions contains many visual examples of the correct output for a given input. Unclear instructions may increase annotator frustration and fatigue so aim to reduce ambiguities.

In addition, make sure that annotators have enough information **readily available** within their tool when completing non-trivial tasks (e.g. in the case of ImageNet classification, the workers were provided with a target concept, like “Burmese cat”, its definition from WordNet, a link to Wikipedia, and a collection of candidate images.)

We also recommend sharing background information with the labelers about the intended goal and use of the model. **Informed labelers are empowered labelers** and they will be able to support you much better if they understand the actual purpose of their labeling task.

Ethics and bias checks would ideally be incorporated into the labeling task as well and labelers can be instructed how to report biases or ethical issues in the data that they come across.

2. Edge case examples

Depending on your use case, there might be a variety of edge cases but some examples to consider are: cropped or truncated objects, occluded objects, very small objects, blurry or distant objects, extreme closeups, groups of objects together, etc. Ensuring that labelers are performing the annotation of these cases in a **consistent way** is key to a clean dataset.

For example, when annotating images of house plants, labelers may encounter images of wild plants for which there is no good class, or drawings of plants which are not actual plants.

If annotators are prepared to handle such cases (e.g. they can use a class **"other"** or flag the image), they can help you spot and eliminate such cases rather than introduce noise into the dataset.

One other strategy is allowing labelers to tag images as **"unsure"** while they are annotating which helps to separate non-prototypal and difficult images from the rest.

Uncertainty ratings (e.g. a scale from 0 to 10) on the image and object level can be applied across the entire dataset, so that annotators record their certainty each time they create an annotation. This can help not only with quality control but also with the evaluation of the model's performance on "difficult" vs "easy" images.

3. Communication & feedback

While there occasionally are malicious or adversarial workers, it's safe to assume that most labelers are motivated to perform well and in good faith. A very constructive approach which is time-consuming but pays off is to establish **channels for communication** and feedback both among workers and between them and the client.

The existence of such channels is a given in managed annotation teams, where special care is taken to train workers, make sure they are all on the same page in their interpretation of the data, and transmit feedback (including having the annotators correct their own mistakes).

Even though crowdsourcing platforms do not usually offer such features, crowdsourced workers have set up external forums where requesters can participate as well.

Iterations & audits

Feedback loops

Ensuring bias-free models requires continuous attention throughout the AI project lifecycle: from framing the problem, collecting and annotating the data, through training the model, evaluating it and deploying it, to auditing the deployed model and fine-tuning it.

At each stage, we recommend starting with **small iterations** and planning how feedback and insights will be collected and propagated to other stages. For example, during labeling it might turn out that the chosen taxonomy does not cover edge cases that are so important or common that they need to be assigned a class of their own.

If this is established after a quick first iteration, it will be much more **cost-effective** than having to re-label the entire dataset. Similarly, if during annotation the distribution of certain classes appears to be unequal, there needs to be a clear process for feeding this back to the collection stage and addressing it in a timely manner.

Once a model has successfully been trained and deployed, its relevance can be ensured by **replenishing its training data often**. The environment in which the model operates is frequently non-stationary and there might be considerable data drift.

Therefore the model needs to be periodically retrained and refreshed using new data sets so as not to become outdated.

Beyond train/val/test

We at Humans in the Loop promote the continuous evaluation of models once they are being deployed on real-life data.

Currently, the most prevalent method of evaluating a model's performance is to split the ground truth dataset into a "training", "validation", and "test" portion.

However, if the entire dataset is biased or unrepresentative of the real world, using one portion of it for testing would give **falsely high results**. Therefore, models should preferably be evaluated based on their performance in real-life practical situations.

For the purposes of model evaluation, professional humans in the loop can be plugged into the workflow. They can regularly validate the proposed labels and perform error analysis by reviewing and classifying the errors.

Error classification following this methodology can be used to distinguish between model errors (localization error, confusion with semantically similar objects, false positives on background, duplicate boxes) or ground truth errors (mis-labeled data: missing labels, incorrect labels or incorrectly grouped objects).

This will help to provide insights into which classes the model performs poorly on, whether it is exhibiting any biases, and whether it makes certain recurring mistakes due to potential systematic errors in the training dataset.

Model monitoring

Model monitoring can be performed in a variety of ways. It's more straightforward in the case of object detection and semantic segmentation than in the vase of image classification. In the first case, a review of the resulting annotations and segmentation masks is enough but in the second one **explainability** is a considerable issue and it requires the use of saliency maps or activation maps. These highlight the regions of each image that contribute most to the model's prediction.

Another option is using a tool like LIME or SHAP which use a superpixel approach to measure how each superpixel affects the prediction. By reviewing the outputs, the human evaluators can judge more easily what went wrong and how to correct it by amending or extending the training dataset.

Model monitoring and dataset auditing are still in a nascent phase but we believe that they can contribute to detecting and mitigating biases in AI systems early on. These can be complemented with **documentation** practices such as Datasheets for Datasets and Dataset Nutrition Labels in order to bring more transparency and accountability across the AI value chain.

We hope the whitepaper series was useful and we look forward to contributing to more AI and computer vision projects that wish to incorporate bias mitigation. Get in touch with us if you would like to discuss how to use better dataset collection and annotation strategies to eliminate bias in your AI solution!

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