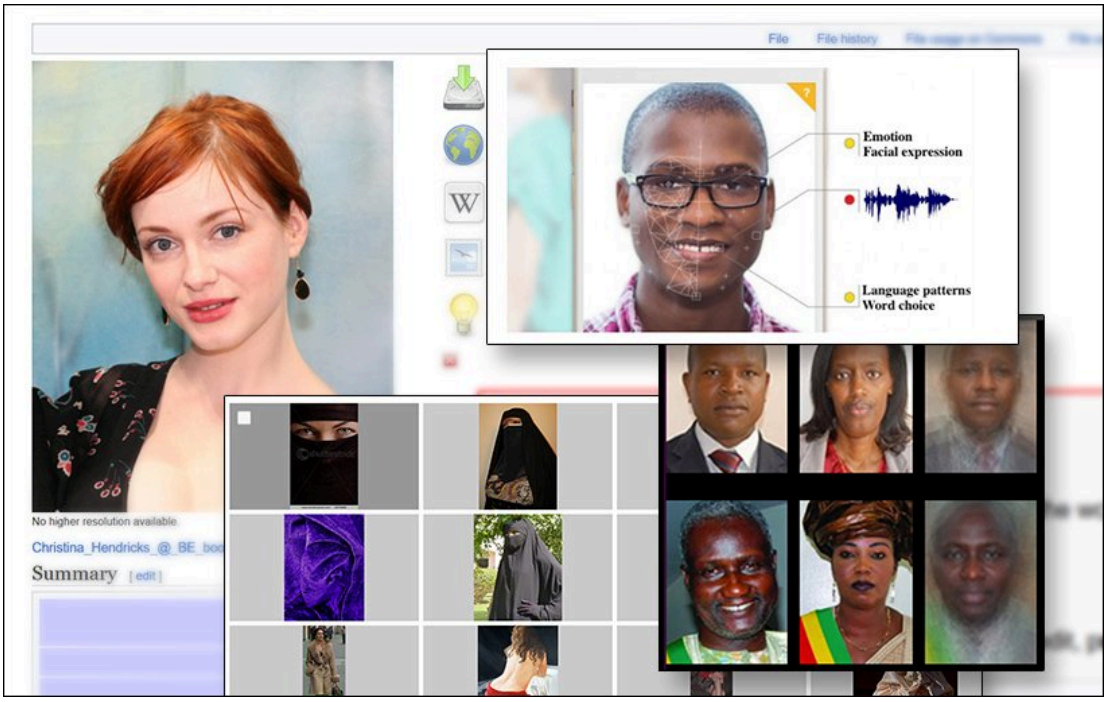
# Dataset Abuse Is Rife in Computer Vision – But the Solutions May Be Drastic

Originally published at https://arquivo.pt/wayback/20240203185557oe_/https://blog.metaphysic.ai/dataset-abuse-is-rife-in-computer-vision-but-the-solutions-may-be-drastic/

February 9, 2023




A new paper, from a cohort of researchers at Sony AI, levels extensive criticism at the *laissez faire* attitude of the machine learning research community towards the ethical use of the computer vision datasets that fuel headline-grabbing applications such as Stable Diffusion.

The new work posits multiple ways in which modern usage – and abuse – of historical datasets is contravening both the letter and the spirit in which they were originally compiled, and offers some radical solutions – albeit at what many may consider a huge burden of cost and friction to the current breakneck pace of development.

## Move Fast and Break Faith

As the paper notes, the community's hunger for hyperscale collections of human-centric images has largely replaced the highly-curated datasets that characterized the sector from the 1970s until more recent times – a period in which there were no easily-available network resources from which to garner billions of images at will; but in which it was much more common that those included in datasets containing their image would have been required to give consent; would have had some idea what line of study their participation was facilitating; and would have had reasonable expectation that their image would not be later used for what they might have considered at the time to be 'unethical' research.

The current scene is practically the inverse of this, in that faces are harvested at scale without the knowledge of those depicted, and either without permission or under licenses (including open source license) which remove agency from the participants, and require them to 'opt out' (if such mechanisms even exist, which they rarely do, even where applicable laws might require it).

The new paper is extensive and comprehensive, so let's take a look at some of the essential problems identified therein, and at some of the remedies envisioned by the researchers.

## 'Legacy' Datasets Repurposed

The researchers observe that many machine learning datasets get used for a purpose other than the one for which they were created, throwing into question the continuing validity of any consent that the participants may have originally given for being included in the work.

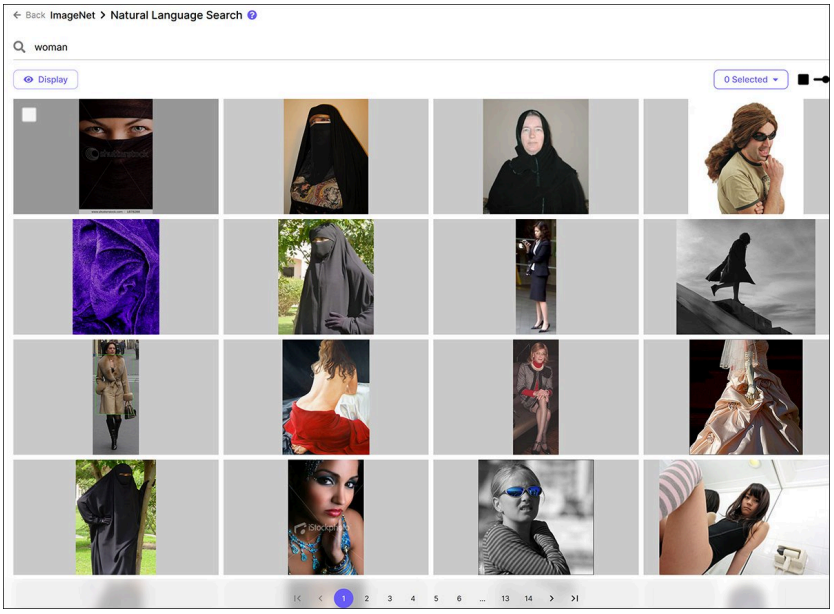
That consent, and any surrounding agreements, may arguably have been granted in an entirely different context, and, in some cases, in an entirely different era, when current possible technological applications for the data may not even have been envisioned.

Citing prior work on the issue from London South Bank University, the paper states*:

*'[When] ML problems are improperly formulated and a dataset is conceived first, the dataset ends up motivating the set of tasks it is used for. This corresponds to situations where a dataset already exists and is repurposed. This can happen when no suitable dataset for a task exists or a dataset's purpose was never clearly delimited.*

*'The latter is often a consequence of inadequate dataset documentation, particularly when dataset creators do not clearly specify a dataset's terms of use. This can raise ethical concerns if a dataset is used for an unintended or dubious application; extended or modified; or, distilled into a model.'*

The evolution of the computer vision research sector is so labyrinthine, intertwined and incestuous that the provenance of all the weights in any individual algorithm is not always easy to establish. Frequently, during model training or dataset pre-processing, libraries are invoked (and this includes some very basic mechanisms such as loss functions) that, somewhere upstream, have been conditioned by 'ancestral' datasets like the venerable ImageNet.




*ImageNet was originally conceived to help research into object recognition, though its application has been widely extended to other sectors, not least of which is the study of human physiognomy – an incidental aspect in the dataset's early days. Source: https://dashboard.scale.com/*

Since there's no established 'chain of custody' for the data that transits and evolves downstream in this way (or at least none that is universally respected), an extraordinary number of datasets tend to end up used in subsequent projects that may have quite a different scope to the ambit of the original project.

The LSBU paper cited earlier notes the 'relaxed' attitude among researchers to the rights and permissions related to datasets:

*'[We] found anecdotal evidence that non-commercial dataset licenses are sometimes ignored in practice. One response reads: "More or less everyone (individuals, companies, etc) operates under the assumption that licences on the use of data do not apply to models trained on that data, because it would be extremely inconvenient if they did."*

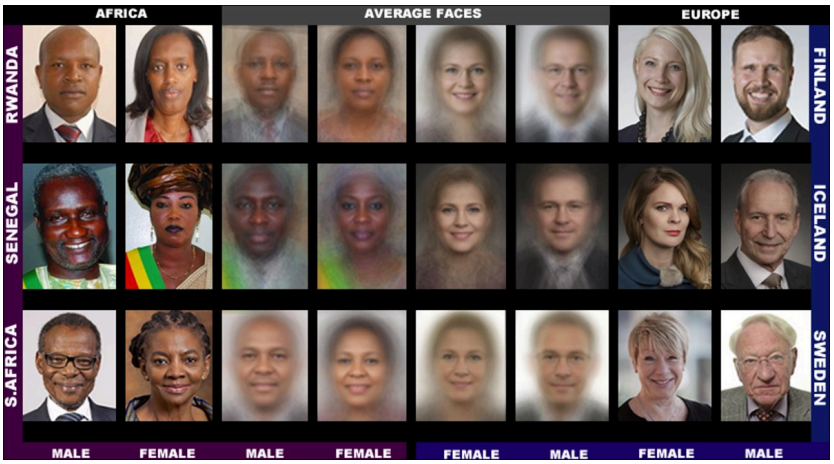
*'Another response reads: "I don't know how legal it really is, but I'm pretty sure that a lot of people develop algorithms that are based on a pretraining on ImageNet and release/sell the models without caring about legal issues."*

*'It's not that easy to prove that a production model has been pretrained on ImageNet …" Commonly-used computer vision frameworks like Keras and PyTorch include models pre-trained on ImageNet, making the barrier for commercial use low.'*

## Fairly Ineffective

Though there are a small number of 'fairness-aware' datasets, the authors criticize these as inadequate, and 'incompatible with common computer vision tasks', having been obtained from the internet, like all the fairness-*un*aware datasets, and lacking the labeling and structure common to such sets, which makes direct comparisons between fair and unfair sets effectively impossible.

The paper opines that besides their relative general scarcity, such fairness-aware datasets as are available – including Pilot Parliaments, AHP and Image Embedding Association Test (iEAT) – contain an inadequate volume of data and too narrow a task scope to accomplish their purpose, with limited labels (though it could be argued that the lack of any kind of consistent inter-dataset labeling schema is a further factor limiting the effectiveness of such efforts).



*The Pilot Parliaments dataset is a 'fairness-aware' initiative. Source: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf*

In any case, 'like-for-like' comparisons can often be hindered by shortcomings in the schema of a typical dataset: the authors note also that many datasets, such as Common Objects in Context (COCO), lack any information about the people depicted in the images, obliging researchers evaluating fairness to use crowdsourced annotators, who in themselves are not above reproach and potential bias.

*'Furthermore,'* they state*. *'such datasets lack ground-truth, self-identified labels about the image subjects, as the information is typically collected indirectly from online resources.'*

The researchers' suggested remedy is pragmatic: *'refrain from repurposing existing datasets'.*

Instead they recommend that datasets should be 'carefully curated', with assiduous metadata, including data related to demographic information, environment, and the specifics of the capture devices used (such as focal length, as well as any software used to produce the image).

## 'Purpose Creep'

They also suggest that datasets should contain 'purpose statements', since subsequent projects that wish to make use of the data could then be evaluated against these (and possibly found wanting).

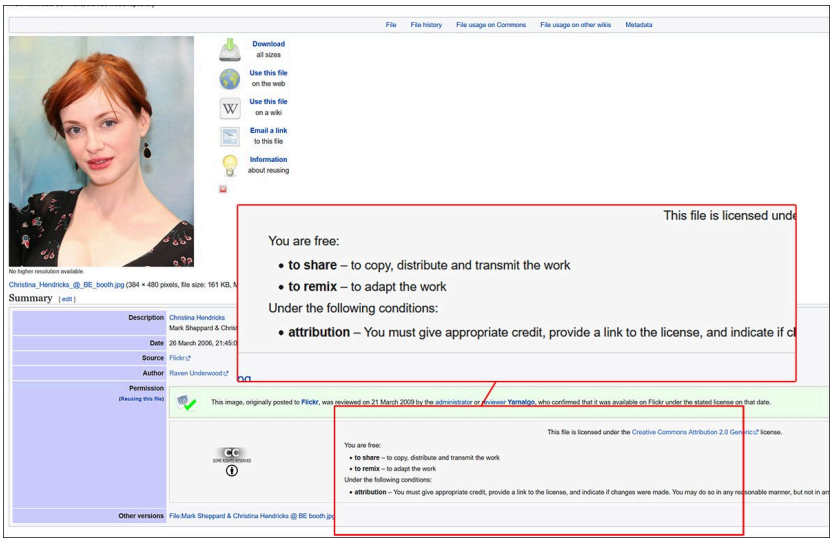
The statements, the authors say, should include what the original objective of the dataset is, who the intended consumers are, and also some information about what the set is *not* to be used for.

*'Dataset documentation can be manipulated to* fit *the narrative of the collected data, as opposed to* directing *the narrative of the data to be collected. Purpose statements which are defined prior to data collection can mitigate against* purpose creep*, i.e., the gradual widening of the scope of a dataset beyond its original purpose.'*

One example the authors give of this is when stakeholders (i.e., the people whose faces and/or bodies appear in image datasets) may approve of the original intent of a dataset, but not wish that their data be used in research intended to improve government surveillance systems.

## Consent and The Creative Commons Loophole

Addressing the issue of consent, the new research observes the way that multiple datasets have been created via the use of the Creative Commons Loophole (CCL). This occurs when people give away rights to data that they do not necessarily own or have the right to dispose of in this way, such as when a photographer 'donates' an image of a person, using a very liberal Creative Commons license.




*One of many cases where the photographer grants the user a liberal license to reuse their work, but without the need for permission from the subject of the work to be included in AI-facing datasets that may not align with their beliefs. Here attribution is required, at least for the photographer, though even this minimal and self-serving requirement is not certain to survive dataset preprocessing. Source: https://commons.wikimedia.org/wiki/File:Christina_Hendricks_@_BE_booth.jpg*

This is broadly accepted because the person originating the image (i.e., pressing the shutter button) is considered to be the owner, and the subject is thus considered to be 'fair game' (depending on jurisdiction, and on any exceptions made for public figures).

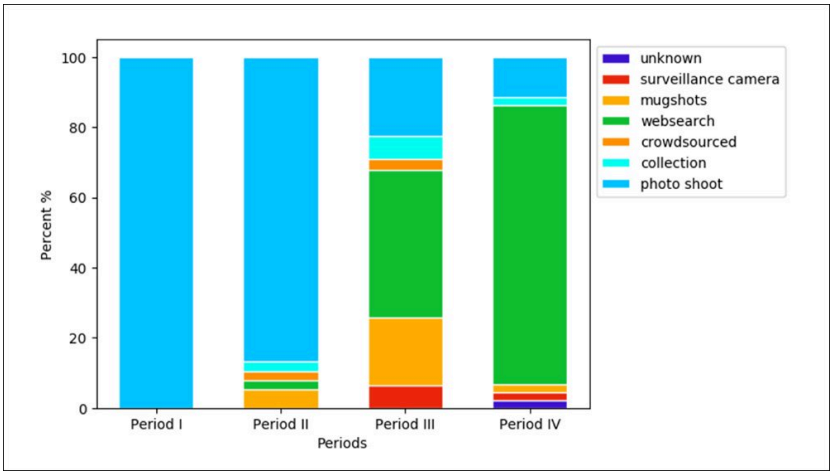
The paper observes that the CCL contravenes a number of state and international regulations, despite its widespread use as a convenient way of skirting consent issues around inclusion in datasets. These include the Illinois Biometric Information Privacy Act (BIPA), article 4(11) of the UK and EU General Data Protection Regulation (GDPR), and article 29 of China's Personal Information Protection Law (PIPL)

The authors note*:

*'Although a Creative Commons license can unlock restrictive copyright, this only pertains to the image regions "that contain copyrightable artistic expressions", rather than image regions used by [computer vision] models containing biometric data such as faces, which are safeguarded by privacy and data protection laws .'*

## Death of the 'Slow' Dataset

The authors discuss also the 'deprecation' of 'slow and considered' data collection practices, where constrained photography was the norm.




*Covering a period from 1976 to 2019, we can see that in the mid-1970s 'manual' and close-to-hand methods of data generation were ubiquitous – a 'slow' and low-volume method that has long since been supplanted by web-scraped online data, and by capture methods that lack control mechanisms, accountability, or context. Source: https://arxiv.org/pdf/2102.00813v1.pdf*

The paper observes, in effect, that originally-created and curated data of this kind is now seen as to some-extent 'unreal' or 'inauthentic' to researchers, who are looking for insights from spontaneous in-the-wild' sources.

Citing a 2020 study, the new work even casts the common use of web-crawling to obtain data in the light of *imperial* practices:

*'A principal source of ethical concern emanates from the shift to, for example, web scraping as a de facto means of unconstrained data collection, regarding people as objects without the right to privacy, or the agency to consent or opt-out.*

*'This is analogous to colonialist attitudes, whereby human image subjects are treated as "raw material free for the taking".'*

Inevitably, the older and more assiduous methods could not possibly scale up to the sheer volumes of data obtainable by high-scale web-crawling, notwithstanding the ethical or legal concerns that the practice may invoke.

## Revoking Consent

The authors observe also the notable difficulty entailed in removing one's image from a dataset, once it's in circulation (by which time, as mentioned above, it may have already filtered downstream into other mechanisms that will not, in any case, necessarily reflect any later revocations of consent):

The paper notes that one of the very few major computer vision datasets to offer such a means to retract is the FFHQ set, which allows users to opt out of inclusion in the collection – a solution the authors consider unsatisfactory*:

*'As image subjects did not consent to being included in the dataset, there is no reason to believe they have any knowledge of their inclusion. This renders the consent revocation offering hollow.*

*'Moreover, these processes place the burden on data subjects to track down uses of their data in datasets, "which are often restricted to approved researchers".'*

It seems that the computer vision research sector is, in this respect, similar to the public's broad (if cynical) perception that once data is online, it's practically impossible to remove, and will be used *ad hoc*, and as anyone who has access to it sees fit – not least because the (far from complete or globally-applicable) laws that impede this make some exception for 'research purposes', regardless of whether such research may later fork into commercial branches, or into more contentious areas than its initial aims.

## Metadata

The new work acknowledges that the use of metadata in images is both part of the problem, but also potentially part of the solution to some of the extensive grievances listed therein.

On the one hand, metadata attached to images, the paper notes, can reveal personally identifiable information (PII), and the authors remind us that some of the most popular targets for image-centric web-scrapers, such as Flickr, assiduously preserve (and sometimes controversially augment) these metatags.

On the other hand, metadata can be a mitigating and ameliorating factor, in that it can provide useful details about the capture device, so that, for instance, the level of lens distortion in an image can be estimated by metadata that contains information about the focal length of the capture device's lens.

Here, the authors' suggestions could potentially raise eyebrows in the machine learning community, though many of the recommendations are only a potential actuation of existing laws (particularly in the EU, under GDPR).

One measure suggested by the researchers is that dataset curators obtain 'voluntary informed consent', in order to use casually-obtained images of people, stating that this safeguards against future litigation or conflict.

*'We recommend that explicit informed consent is obtained from each person depicted in, or otherwise identifiable by, the dataset using plain language consent and notice forms. In particular, data subjects should have voluntarily consented to the sharing of their data, including their facial, body, biometric, and other images and information about themselves and their surroundings, for the purposes of developing, training, and/or evaluating CV technologies.'*

The authors further recommend that inclusion of material relating to minors require an additional layer of consent.

It can be argued that the adoption and enforcement of this practice would effectively lay waste to the growing ecostructure of hyperscale datasets that are powering the current generation of latent diffusion models, and would have a substantial impact on the development of new computer vision technologies in general – particularly if older, non-compliant datasets were not allowed to be ring-fenced from the effect of such an edict.

## Further Considerations

The paper's examination of the issues around dataset legality and ethics are, as noted, comprehensive, and we invite readers to take a deeper dive into the source material. Some of the other issues covered are worth briefly noting, however.

In regard to representational bias, the authors note the many instances in which both people of color, and older people, have been historically misrepresented or discounted by algorithms developed from biased or poorly-labeled data.

To address this, the authors recommend *self-reported annotations*, where the subject actively contributes to labels and classes that define their own data. Arguably, this is an even more demanding schema than requiring consent, since it involves active participation and involvement by the participants, and a level of attention to detail that's more commonly associated with small volunteer student studies from the 1970s and 1980s.

Further, though there would apparently be no arguing with the authenticity and provenance of user-generated annotations, their objectivity could, some might argue, be subject to question.

However, the paper notes that this kind of user-created information could be the only way to define marginal or edge cases in categorization, such as cases where the user may not identify with their perceived gender, or where the user's sex and gender may not be clear from images that contain them.

In regard to whether an image can be used according to regulations that apply to the person depicted, the authors suggest that researchers collect *country of usage* information.

*'This will help.'* the authors state*, *'to ensure that each subject's data is protected accordingly and will assist dataset creators in appropriately addressing future legislative changes. For example, GDPR Article 7(3) explicitly states that data subjects have the right to withdraw their consent*

*at any time, which was not explicitly addressed under its predecessor.'*

The paper also suggests, as recent research has done, that greater attention be paid to the qualities of the people annotating the data, who may be operating under constraints and circumstances that are non-optimal, and who themselves are likely to bring some level of bias to the gathered data.

## Conclusion

In summary, the proposals offered by the Sony AI researchers could be regarded as both visionary and revolutionary, but also, to a certain extent, atavistic: the changes envisaged constitute, in effect a return to the higher standards and practices of the 1970s, when computer vision was a fringe and underpowered pursuit in the academic scene.

In regard to the effect that the proposed changes would have on the multi-billion dollar applied computer vision sector, and on the extent to which governmental anxiety is fueling the current pace of progress, the researchers appear to offer no solutions or palliative compromises.

** My substitution of the researchers' inline citations for hyperlinks.*