Humans Categorise Humans: on *ImageNet Roulette* and Machine Vision

Olga Goriunova

Published in Donaufestival: Redefining Arts Catalogue, April 2020.

In September 2019, artist Trevor Paglen and researcher Kate Crawford were getting through the final week of their *ImageNet Roulette* being publicly available on the web. *ImageNet Roulette* is part of their exhibition "Training Humans," which ran between September 2019 and February 2020 at the Foundazione Prada. With this exhibition, collaborators Crawford and Paglen queried the collections of images and the processes used to train a wide range of machine learning algorithms to recognise and label images (known as "machine vision").[1] Crawford and Paglen want to draw attention to how computers see and categorize the world: seeing and recognizing is not neutral for humans (one need only think about "seeing" race or gender), and the same goes for machines.

Machine vision, by now a large and successful part of AI, has only taken off in the last decade. For "machines" to recognize images, they need to be "trained" to do so. The first major step for this to happen is to have a freely available and pre-labelled very large image-data-set. The first and largest training dataset that Crawford and Paglen focused on and which they derived the title of their project from is ImageNet, which is only ten years old.

ImageNet Roulette is a website or an app that allows one to take a selfie and run it through ImageNet. It uses an "open-source Caffe deep-learning framework … trained on the images and labels in the 'person' categories."[2] One is "seen" and "recognized" or labelled. I took my picture sitting with my laptop in my bed, in pyjamas and with bad lighting, and was labelled "person, individual, someone, somebody, mortal, soul => unwelcome person, persona non grata => disagreeable person => creep, weirdo, weirdee, wierdy, spook." So far, so good: I don't mind being labelled weird: perhaps it's part of my intellectual appeal. I ran it again in the daytime, with the same result. It quickly became clear though, that the labelling carried substantially more serious consequences for people of color. *Guardian* journalist Julia Carrie Wong was labelled "gook, slant-eye." She wrote: "below the photo, my label was helpfully defined as "a disparaging term for an Asian person (especially for North Vietnamese soldiers in the Vietnam War),"[3] My younger female southern European PhD student told me she was labelled a "virgin," and when she ran the app again, a "mulatto," but her British male Black friend was labelled simply a "rapist."

This verdict of ImageNet Roulette is indicative of bias and an in-built racism, the evidence of which accompanies the developments in machine vision and image recognition technologies. In 2009, an HP face-tracking webcam could

[1] For the research paper, see: Kate Crawford and Trevor Paglen, "Excavating AI: The Politics of Images in Machine Learning Training Sets"; 19 September, 2019;

[2] Ibid.

[3] Julia Carrie Wong, "The viral selfie app ImageNet Roulette seemed fun—until it called me a racist slur," *The Guardian*, 18 September 2019, https://www.theguardian.com/technology/2019/sep/17/imagenet-roulette-asian-racist-slur-selfie

not follow a Black person.[4] In 2010, Microsoft's Kinect motion-sensing camera did not work so well with dark-skinned users.[5] In 2015, Google Photos classified Black men as gorillas (a "feature" Google fixed by removing "gorilla" from its set of image labels).[6] Joy Buolamwini, a Black researcher, had to wear a white mask to test her own software.[7] Face recognition, action recognition, and image categorization are different technical procedures whose results converge when it comes to privileging white skin and Caucasian features. It is not dissimilar to discrimination in other fields: voice-command car systems often do not respond to female drivers, but are quick to obey male voices, even when they come from the passenger seat.[8] How do machines pick up racism and sexism? How do they learn to be nasty?

By now, there are some standard responses to these questions. Machine learning is a branch of computer science that develops algorithms that improve their performance with experience. Neural networks are one example of machine learning algorithms, which are said to be modeled on the human neural network and capable of handling complexity. They need to be trained, either by a human, or in an automated way, on a dataset; after training they are able to start making independent decisions. For instance, one runs a neural network through a dataset of many images of haircuts, teaching the network to recognize different types of haircuts. This is done by saying repeatedly, for example, "this is a bob," and "this is a shaved head." The neural network can then start to recognize haircuts on its own. What matters here are the model and the dataset. The model might weigh certain qualities of hair and kinds of haircuts higher than others, making it, for instance, unable to differentiate between types of haircuts or, say, curly hairstyles. In this way, the training dataset does not become diverse enough; or it is annotated in such a way that makes it biased towards certain kinds of hair. If curly hair is not included in the dataset, or included with negative labels, the new outputs would sustain the "bias." Beyond models and datasets, computer vision's infrastructures—which include server architectures, the labor of computer scientists and data workers, knowledge infrastructures that spread beyond one discipline and format, and many other elements and processes—can all carry "bias."

As *ImageNet Roulette* made headlines around the world, The Photographers' Gallery in London—and in particular the curator of Digital Programmes Katrina Sluis, a long standing scholar of the transformations that algorithms and AI bring to the field of photography and visual culture, and researcher Nicolas Malevé—organized a symposium on computational images (as well as a birthday party for ImageNet), securing the participation of Fei-Fei Li, Professor at the University of Stanford, and one of the creators of ImageNet. Li's lecture, celebratory and focused on the history and effort that went into ImageNet, was, in its framing, orientation points and identification of inspiration, illuminating.

[4] "HP looking into claim webcams can't see black people." CNN, 24 December 2009; http://edition.cnn.com/2009/TECH/12/22/hp.webcams/index.html

[5] "Is Microsoft's Kinect racist?" *PCWorld*, 4 November 2010. https://www.pcworld.com/article/209708/Is_Microsoft_Kinect_Racist.html

[6] Tom Simonite, "When it comes to gorillas, Google Photos remains blind," *Wired*, 11 January 2018; https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

[7] Joy Buolamwini, "InCoding—in the beginning," *Medium*, 16 May 2016; https://medium.com/mit-media-lab/incoding-in-the-beginning-4e2a5c51a45d

[8] Sharon Silky Carty, "Many Cars Tone Deaf To Women's Voices," Autoblog.com, 31 May 2011, https://www.autoblog.com/2011/05/31/women-voice-command-systems

ImageNet is an image dataset that helped make a breakthrough in computer vision. Without a large dataset of images, no major work on the automation of vision would be possible. Fei-Fei Li started working on ImageNet in 2009, as she recounted, against advice from senior colleagues. ImageNet's images come from Flickr, and were automatically harvested by the millions. No computer vision breakthrough would have been possible without social media and mass uploads of user-generated images, free to use. At first, Princeton undergraduate students were paid to categorize and label images. It was expensive and slow work. At the same time, Amazon's "Mechanical Turk" was launched. What has become known as the marketplace for outsourcing the tedious and painstaking labor undergirding digital culture to countries where people have to accept earning as little as 0.02 USD per task, was also key to the success of machine vision. 50,000 workers in over a hundred countries undertook the labor of sifting through 160 million candidate Flickr-derived images, and annotating 14 millions of them, by using the WordNet semantic structure.[9] WordNet is a lexical database developed from 1985 onwards and used for automated text analysis (machine translation, information retrieval, and other tasks). The database works as a "conceptual dictionary," grouping words into sets of synonyms (synsets), giving short definitions and examples; it further organizes them into hierarchies, going from the specific to the more abstract. WordNet is notorious for its bias: query it for "woman" and you get the following.[10]

- S: (n) **woman**, adult female (an adult female person (as opposed to a man)) *"the woman kept house while the man hunted"*
- S: (n) **woman** (a female person who plays a significant role (wife or mistress or girlfriend) in the life of a particular man) *"he was faithful to his woman"*
- S: (n) charwoman, char, cleaning woman, cleaning lady, **woman** (a human female employed to do housework) *"the char will clean the carpet"; "I have a woman who comes in four hours a day while I write"*
- S: (n) womanhood, **woman**, fair sex (women as a class) *"it's an insult to American womanhood"; "woman is the glory of creation"; "the fair sex gathered on the veranda"*

Needless to say, the entry on "man" is three times longer and does not define the man exclusively in relation to sexual and household services to the "opposite sex." Once WordNet's semantic structure was taken on as the system to label images, it was up to the workers sourced through the Mechanical Turk to decide which categories to "image." As Fei-Fei Li explained, the process was entirely automated. "Love" seemed not "imageable," but up to 2,833 "person" categories were deemed to be. As Crawford and Paglen write, "The subcategory with the most associated pictures is "gal" (with 1,664 images), followed by 'grandfather' (1,662), 'dad' (1,643), and chief executive officer (1,614). With these highly populated categories, we can already begin to see the outlines of a world view. ImageNet classifies people into a broad range of types including race,

[9] The process of annotation itself, including research and tests it is based on, timing, and the process of eliminating inaccurate annotators, demands a separate investigation. See the work of Nicolas Malevé, including his forthcoming PhD thesis. Some of his work is available here: https://unthinking.photography/contributors/nicolas-maleve

[10] WordNet is accessible through the browser: http://wordnetweb.princeton.edu/perl/webwn

nationality, profession, economic status, behavior, character, and even morality."[11]

The question is, where is the racism exhibited by ImageRoulette located? Is it in the WordNet that includes inappropriate and offensive categories and words? Is it in the workers who chose those "person" categories? Is it in the 100+ countries whose cultures informed the labeling of the images? Is it enhanced by the system of the Mechanical Turk? Is it now inscribed into the ImageNet dataset? Is it something the creators of ImageNet must have considered and tried to counter-act at the point of design? As a response to *ImageNet Roulette*, work has started on "debiasing" ImageNet, by the annotation of synsets in terms of their "offensiveness," resulting in the removal of 600,004 images, by addressing "imageability" of concepts, and by taking measures to diversify images.[12]

There is still more to how computer vision has developed. From the way Fei-Fei Li presented her field in her lecture, it was very clear that computer vision is squarely modeled on human vision. While one could think that human vision, while certainly complex, is fairly limited, and that the visual capacities of other animals, including insects, can be seen as far superior, this is not what the computer vision field seems to believe. Putting the human firmly in the center is very problematic. After all, isn't the belief that the human is the acme of evolution (or as previously said, of creation) and a universal measure of things the reason why the world is facing a climate catastrophe and the sixth mass extinction is under way? A certain quasi-religious and Renaissance-inspired appreciation of the human is propped up by capitalism. The human is, as Shoshana Magnet has put it, the source of all possible cash, and capital remains the driving force *par excellence* of computer vision.

Fei-Fei Li recounted that one of the biggest breakthroughs in computer vision came from cognitive neuroscience and its discovery of the part of the human brain—the inferior temporal cortex—that is responsible for visual object recognition. By the age of five, a human child learns to recognize thousands of different objects, including highly abstract shapes. Human visual perception of objects, as well as language and other human qualities, are here considered more complex and far superior to anything else in the living world—a view unproblematically adopted by computer scientists. Intelligence in this case is not rational thought, but visual object recognition modeled on the human brain or the complexity of language, specifically construed in human terms so as to discount intelligence in whales or in gorillas. In machine vision, it is the human visual perception of objects that the machine is set to approximate, and even to replicate. It is based on the human, and for the human. Since the Turing test (if a bot chatting to a human for a set amount of time can convince them that they are chatting to a human rather than a bot, the test is passed), which redefined machine intelligence as the capacity to appear to humans as intelligent, rather than to be, in itself, intelligent, the human measure has been firmly set at the heart of AI: "Can humans detect objects? Then let's detect objects too!" The horizon of the next developments in AI is constructed in the same terms: "Do humans interact with the world? Then let's get the agent moving!"

[11] Crawford and Paglen, 2019.

[12] Kaiyu Yang (Princeton University), Klint Qinami (Princeton University), Li Fei-Fei (Stanford University), Jia Deng (Princeton University), Olga Russakovsky (Princeton University), "Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy," 17 September, 2019; http://image-net.org/update-sep-17-2019.php

Furthermore, ImageNet images are common everyday Flickr photos. They are not of great quality and are certainly not outstanding in their aesthetics. They are far from any radical avant-garde photography. What is striking about the origin of images in ImageNet, is how far removed they are from any idea of the autonomy of the machine. The pictures are taken by humans, and from angles seen by or arranged to be seen by humans. For instance, the pictures of birds on display during Fei-Fei Li's talk were taken from "in profile," instead of from the bottom or the top (how other animals might see them), or from all possible angles (how a super-rational and autonomous all-seeing machine would see the world). Although Fei-Fei Li said that these kinds of pictures are chosen for being "as realistic a representation of the world as possible," she recognised how problematical such a statement sounded in The Photographers' Gallery. Whose realism was this? If the avant-garde in film or photography developed technology-specific vision, and where the artistic achievements of Sergei Eisenstein or Alexander Rodchenko were to liberate the camera from the conventions of nineteenth century painting and let the technology inform the layout of the new visual culture, nothing of this spirit is detectable in computer vision. It is about fitting image to word, not the creation or identification of uniqueness.

Crawford and Paglen emphasise that machine vision is built on "epistemological and metaphysical assumptions about the nature of images, labels, categorization and representation,"[13] But the problem is larger than that. The starting position, the point-of-view, is that of the human; the image-making conventions are those of the end of the twentieth century. They are at least 25 years out of date, if not 125.

Oh, if only we really were excited about how radically different from the human machines can be! They might see things no animal can see! They might use a logic alien to the human! They might go somewhere the human has never imagined! Instead, what we have is a replication of the inadequacy and limitations of the human, a vision circumscribed by the horizon of the human.

Certainly, what the industry calls "the bias"—a problematic term inherited from psychology—would manifest itself when the labor of labeling is performed by biased humans. According to psychology, there is no such thing as being unbiased. The only remedy is to become aware of one's biases and try to remedy them by various procedures: for instance, making decisions in groups comprised by a diversity of people; taking time to make decisions; and eliminating stress. None of these aims are met by the conditions in which labeling takes place in computer vision. Annotators take fractions of seconds to look at images, they do it alone and under pressure to perform at speed to earn their meager financial reward. Computer scientists developing and adjusting models do not have the training in or an understanding of the social and political significance of the decisions that they make. But above all, the entire field seems to be modeled in a way that has bias at its core—because it has the notion and the measure of the human at its core.

In that, machines are not at all alien to us. Machines are modeled on us, programmed to replicate and mimic us, and work squarely within our confines of normativity. Feminist, postcolonial, critical, race and disability studies scholars have long taken issue with the words "us" and "humans." "Us" is much differentiated. It is usually a Western white, able man, assuming the capacity to talk for everyone, but in fact excluding people of color, the disabled and women, and framing them as "other"

[13] Crawford and Paglen, 2019.

to "us." Similarly, "human" just masks a differentiation that was historically applied to people that were considered not to have reached the status of the "human," such as first nations and slaves. "Human" masks differences in its claim to modernity, universality, and objectivity.[14] If "human" or "us" is at the center of a project, then there is no doubt that multiple "non-us's" will be discovered, and always predictably along the same lines of race, gender, class, and ability.

Cyberfeminists argued that machines, framed as subservient to the human, as objects and tools, were not unlike women historically forced to be in subservient positions to men, because the generic human always ends up signifying the dominant subject, that is, men.[15] If there is no such thing as "us," machines modeled on "us" would simply reproduce the existing lines of discrimination, tension and struggle. At the moment, digital media have become boring: they are predictable and traditional propaganda, discrimination, and repression machines at the service of capital and power. The question now is whether the machine can actually be liberated from this logic and achieve something of its own capacity, where it is not a servant, but an entity with its own logic of development. A servant always remains inscribed into the violence of a slave-master dialectic. A machine with its own capacities and limitations rather than one that is delimited by the same old "human" problems, would be so much more interesting. A machine that sees what humans cannot see and in a way that does not play into the hands of power. I am looking forward to that machine.

[14] For a discussion of the problems of the category of the human, and possible solutions to these problems, see Rosi Braidotti, *The Posthuman*. London: Polity Press, 2013.

[15] VNS Matrix, "The Cyberfeminist Manifesto for the 21st Century", 1991; https://vnsmatrix.net/projects/the-cyberfeminist-manifesto-for-the-21st-century