Excavating AI: The Politics of Images in Machine Learning Training Sets

Kate Crawford and Trevor Paglen



You open up a database of pictures used to train artificial intelligence systems. At first, things seem straightforward. You're met with thousands of images: apples and oranges, birds, dogs, horses, mountains, clouds, houses and street signs. But as you probe further into the dataset, people begin to appear: cheerleaders, scuba divers, welders, Boy Scouts, fire walkers and flower girls. Things get strange: a photograph of a woman smiling in a bikini is labelled a 'slattern, slut, slovenly woman, trollop'. A young man drinking beer is categorized as an 'alcoholic, alky, dipsomaniac, boozer, lush, soaker, souse'. A child wearing sunglasses is classified as a 'failure, loser, non-starter, unsuccessful person'. You're looking at the 'person' category in a dataset called ImageNet, one of the most widely used training sets for machine learning.

Something is wrong with this picture.

Where did these images come from? Why were the people in the photos labelled this way? What sorts of politics are at work when pictures are paired with labels, and what are the implications when they are used to train technical systems?

In short, how did we get here?

There's an urban legend about the early days of machine vision, the subfield of artificial intelligence (AI) concerned with teaching machines to detect and interpret images. In 1966, Marvin Minsky was a young professor at MIT, making a name for himself in the emerging field of artificial intelligence.[1] Deciding that the ability to interpret images was a core feature of intelligence, Minsky turned to an undergraduate student, Gerald Sussman, and asked him to 'spend the summer linking a camera to a computer and getting the computer to describe what it saw', [2] This became the Summer Vision Project. [3] Needless to say, the project of getting computers to 'see' was much harder than anyone expected, and would take a lot longer than a single summer.

The story we've been told goes like this: brilliant men worked for decades on the problem of computer vision, proceeding in fits and starts, until the turn to probabilistic modelling and learning techniques in the 1990s accelerated progress. This led to the current moment, in which challenges such as object detection and facial recognition have been largely solved. [4] This arc of inevitability recurs in many AI narratives, where it is assumed that ongoing technical improvements will resolve all problems and limitations.

But what if the opposite is true? What if the challenge of getting computers to 'describe what they see' will always be a problem? In this essay, we will explore why the automated interpretation of images is an inherently social and political project, rather than a purely technical one. Understanding the politics within AI systems matters more than ever, as they are quickly moving into the architecture of social institutions: deciding whom to interview for a job, which students are paying attention in class, which

suspects to arrest, and much else.

For the last two years, we have been studying the underlying logic of how images are used to train AI systems to 'see' the world. We have looked at hundreds of collections of images used in artificial intelligence, from the first experiments with facial recognition in the early 1960s to contemporary training sets containing millions of images. Methodologically, we could call this project an *archeology of datasets*: we have been digging through the material layers, cataloguing the principles and values by which something was constructed, and analysing what normative patterns of life were assumed, supported and reproduced. By excavating the construction of these training sets and their underlying structures, many unquestioned assumptions are revealed. These assumptions inform the way AI systems work - and fail - to this day.

This essay begins with a deceptively simple question: what work do images do in AI systems? What are computers meant to recognise in an image and what is misrecognised or even completely invisible? Next, we look at the method for introducing images into computer systems and look at how taxonomies order the foundational concepts that will become intelligible to a computer system. Then we turn to the question of labelling: how do humans tell computers which words will relate to a given image? And what is at stake in the way AI systems use these labels to classify humans, including by race, gender, emotions, ability, sexuality, and personality? Finally, we turn to the purposes that computer vision is meant to serve in our society - the judgments, choices, and consequences of providing computers with these capacities.

Training AI

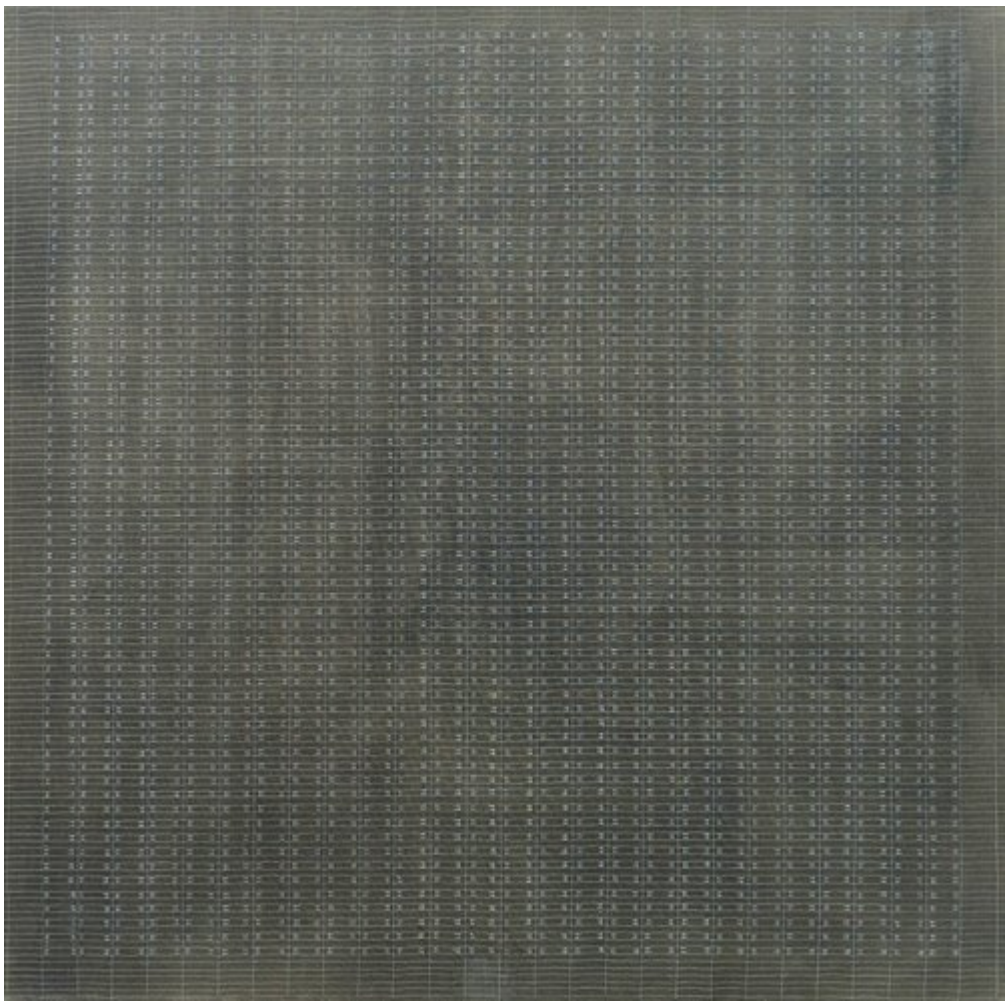
Building AI systems requires data. Supervised machine-learning systems designed for object or facial recognition are trained on vast amounts of data contained within datasets made up of many discrete images. To build a computer vision system that can, for example, recognise the difference between pictures of apples and oranges, a developer has to collect, label and train a neural network on thousands of labelled images of apples and oranges. On the software side, the algorithms conduct a statistical survey of the images, and develop a model to recognise the difference between the two 'classes.' If all goes according to plan, the trained model will be able to distinguish the difference between images of apples and oranges that it has never encountered before.



Training sets, then, are the foundation on which contemporary machine- learning systems are built.[5] They are central to how AI systems recognize and interpret the world. These datasets shape the epistemic boundaries governing how AI systems operate, and thus are an essential part of understanding socially significant questions about AI.

But when we look at the training images widely used in computer-vision systems, we find a bedrock composed of shaky and skewed assumptions. For reasons that are rarely discussed within the field of computer vision, and despite all that institutions like MIT and companies like Google and Facebook have

done, the project of interpreting images is a profoundly complex and relational endeavour. Images are remarkably slippery things, laden with multiple potential meanings, irresolvable questions, and contradictions. Entire subfields of philosophy, art history, and media theory are dedicated to teasing out all the nuances of the unstable relationship between images and meanings.[6]



"White Flower" Agnes Martin, 1960

Images do not describe themselves. This is a feature that artists have explored for centuries. Agnes Martin creates a grid-like painting and dubs it *White Flower*, Magritte paints a picture of an apple with the words 'This is not an apple'. We see those images differently when we see how they're labelled. The circuit between image, label and referent is flexible and can be reconstructed in any number of ways to do different kinds of work. What's more, those circuits can change over time as the cultural context of an image shifts, and can mean different things depending on who looks, and where they are located. Images are open to interpretation and reinterpretation.

This is part of the reason why the tasks of object recognition and classification are more complex than Minksy - and many of those who have come since - initially imagined.

Despite the common mythos that AI and the data it draws on are objectively and scientifically classifying the world, everywhere there is politics, ideology, prejudices and all of the subjective stuff of history. When we survey the most widely used training sets, we find that this is the rule rather than the exception.

Anatomy of a Training Set

Although there can be considerable variation in the purposes and architectures of different training sets, they share some common properties. At their core, training sets for imaging systems consist of a

collection of images that have been labelled in various ways and sorted into categories. As such, we can describe their overall architecture as generally consisting of three layers: the overall taxonomy (the aggregate of classes and their hierarchical nesting, if applicable), the individual classes (the singular categories that images are organised into, e.g., 'apple'), and each individually labelled image (i.e., an individual picture that has been labelled an apple). Our contention is that every layer of a given training set's architecture is infused with politics.

Take the case of a dataset like the 'The Japanese Female Facial Expression (JAFFE) Database', developed by Michael Lyons, Miyuki Kamachi and Jiro Gyoba in 1998, and widely used in affective computing research and development. The dataset contains photographs of ten Japanese female models making seven facial expressions that are meant to correlate with seven basic emotional states.[7] (The intended purpose of the dataset is to help machine-learning systems recognise and label these emotions for newly captured, unlabelled images). The implicit, top-level taxonomy here is something like 'facial expressions depicting the emotions of Japanese women'.

If we go down a level from taxonomy, we arrive at the level of the class. In the case of JAFFE, those classes are happiness, sadness, surprise, disgust, fear, anger and neutral. These categories become the organising buckets into which all of the individual images are stored. In a database used in facial recognition, as another example, the classes might correspond to the names of the individuals whose faces are in the dataset. In a dataset designed for object recognition, those classes correspond to things like apples and oranges. They are the distinct concepts used to order the underlying images.

At the most granular level of a training set's architecture, we find the individual labelled image: be it a face labelled as indicating an emotional state; a specific person; or a specific object, among many examples. For JAFFE, this is where you can find an individual woman grimacing, smiling or looking surprised.

There are several implicit assertions in the JAFFE set. First there's the taxonomy itself: that 'emotions' is a valid set of visual concepts. Then there's a string of additional assumptions: that the concepts within 'emotions' can be applied to photographs of people's faces (specifically Japanese women); that there are six emotions plus a neutral state; that there is a fixed relationship between a person's facial expression and her true emotional state; and that this relationship between the face and the emotion is consistent, measurable, and uniform across the women in the photographs.

At the level of the class, we find assumptions such as 'there is such a thing as a "neutral" facial expression' and 'the significant six emotional states are happy, sad, angry, disgusted, afraid, surprised'.[8] At the level of labelled image, there are other implicit assumptions such as 'this particular photograph depicts a woman with an "angry" facial expression', rather than, for example, the fact that this is an image of a woman mimicking an angry expression. These, of course, are all 'performed' expressions - not relating to any interior state, but acted out in a laboratory setting. Every one of the implicit claims made at each level is, at best, open to question, and some are deeply contested.[9]

The JAFFE training set is relatively modest as far as contemporary training sets go. It was created before the advent of social media, before developers were able to scrape images from the internet at scale, and before piecemeal online labour platforms like Amazon Mechanical Turk allowed researchers and corporations to conduct the formidable task of labeling huge quantities of photographs. As training sets grew in scale and scope, so did the complexities, ideologies, semiologies and politics from which they are constituted. To see this at work, let's turn to the most iconic training set of all, ImageNet.

The Canonical Training Set: ImageNet

One of the most significant training sets in the history of AI so far is ImageNet, which is now celebrating its tenth anniversary. First presented as a research poster in 2009, ImageNet is a dataset of extraordinary scope and ambition. In the words of its cocreator, Stanford Professor Fei-Fei Li, the idea behind ImageNet was to 'map out the entire world of objects'.[10] Over several years of development,

ImageNet grew enormous: the development team scraped a collection of many millions of images from the internet and briefly became the world's largest academic user of Amazon's Mechanical Turk, using an army of piecemeal workers to sort an average of 50 images per minute into thousands of categories.[11] When it was finished, ImageNet consisted of over 14 million labelled images organised into more than 20, 000 categories. For a decade, it has been the colossus of object recognition for machine learning and a powerfully important benchmark for the field.



Interface used by Amazon Turk Workers to label pictures in ImageNet

Navigating ImageNet's labyrinthine structure is like taking a stroll through Borges's infinite library. It is vast and filled with all sorts of curiosities. There are categories for apples, apple aphids, apple butter, apple dumplings, apple geraniums, apple jelly, apple juice, apple maggots, apple rust, apple trees, apple turnovers, apple carts, applejack, and applesauce. There are pictures of hot lines, hot pants, hot plates, hot pots, hot rods, hot sauce, hot springs, hot toddies, hot tubs, hot- air balloons, hot fudge sauce, and hot water bottles.

ImageNet quickly became a critical asset for computer-vision research. It became the basis for an annual competition where labs around the world would try to outperform each other by pitting their algorithms against the training set and seeing which one could most accurately label a subset of images. In 2012, a team from the University of Toronto used a Convolutional Neural Network to handily win the top prize, bringing new attention to this technique. That moment is widely considered a turning point in the development of contemporary AI.[12] The final year of the ImageNet competition was 2017, and accuracy in classifying objects in the limited subset had risen from 71.8% to 97.3%. That subset did not include the 'Person' category, for reasons that will soon become obvious.

Taxonomy

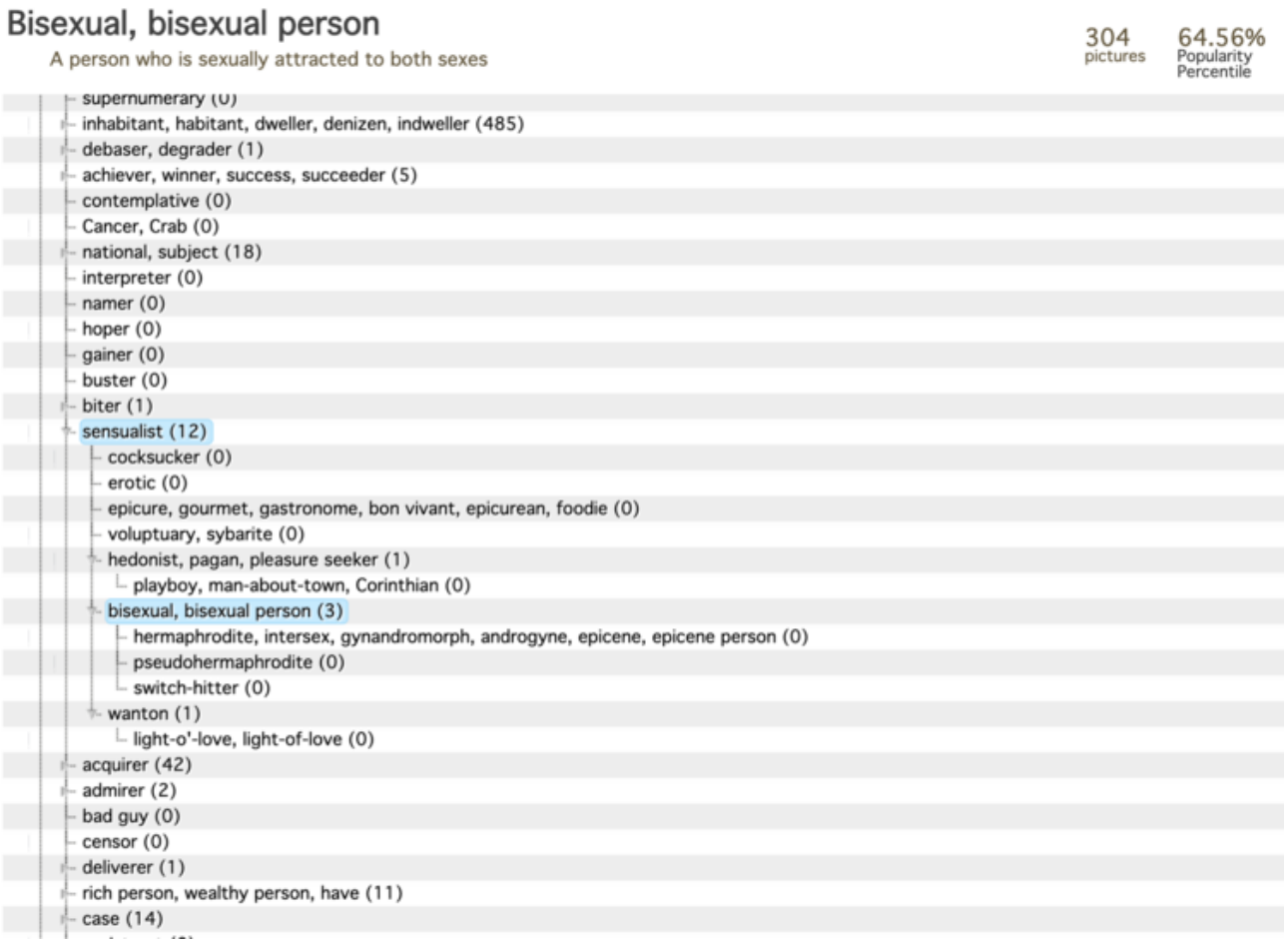
The underlying structure of ImageNet is based on the semantic structure of WordNet, a database of word classifications developed at Princeton University in the 1980s. The taxonomy is organised according to a nested structure of cognitive synonyms or 'synset'. Each 'synset' represents a distinct concept, with synonyms grouped together (for example, 'auto' and 'car' are treated as belonging to the same synset). Those synsets are then organised into a nested hierarchy, going from general concepts to more specific ones. For example, the concept 'chair' is nested as artifact > furnishing > furniture > seat > chair. The

classification system is broadly similar to those used in libraries to order books into increasingly specific categories.

While WordNet attempts to organize the entire English language,[13] ImageNet is restricted to nouns (the idea being that nouns are things that pictures can represent). In the ImageNet hierarchy, every concept is organised under one of nine top-level categories: plant, geologic formation, natural object, sport, artifact, fungus, person, animal and miscellaneous. Below these are layers of additional nested classes.

As the fields of information science and science and technology studies have long shown, all taxonomies or classificatory systems are political.[14] In ImageNet (inherited from WordNet),

for example, the category 'human body' falls under the branch Natural Object > Body > Human Body. Its subcategories include 'male body'; 'person'; 'juvenile body'; 'adult body'; and 'female body'. The 'adult body' category contains the subclasses 'adult female body' and 'adult male body'. We find an implicit assumption here: only 'male' and 'female' bodies are 'natural'. There is an ImageNet category for the term 'Hermaphrodite' that is bizarrely (and offensively) situated within the branch Person > Sensualist > Bisexual > alongside the categories 'Pseudohermaphrodite' and 'Switch Hitter'.[15] The ImageNet classification hierarchy recalls the old Library of Congress classification of LGBTQ-themed books under the category 'Abnormal Sexual Relations, Including Sexual Crimes', which the American Library Association's Task Force on Gay Liberation finally convinced the Library of Congress to change in 1972 after a sustained campaign.[16]



If we move from taxonomy down a level, to the 21,841 categories in the ImageNet hierarchy, we see another kind of politics emerge.

Categories

There's a kind of sorcery that goes into the creation of categories. To create a category or to name

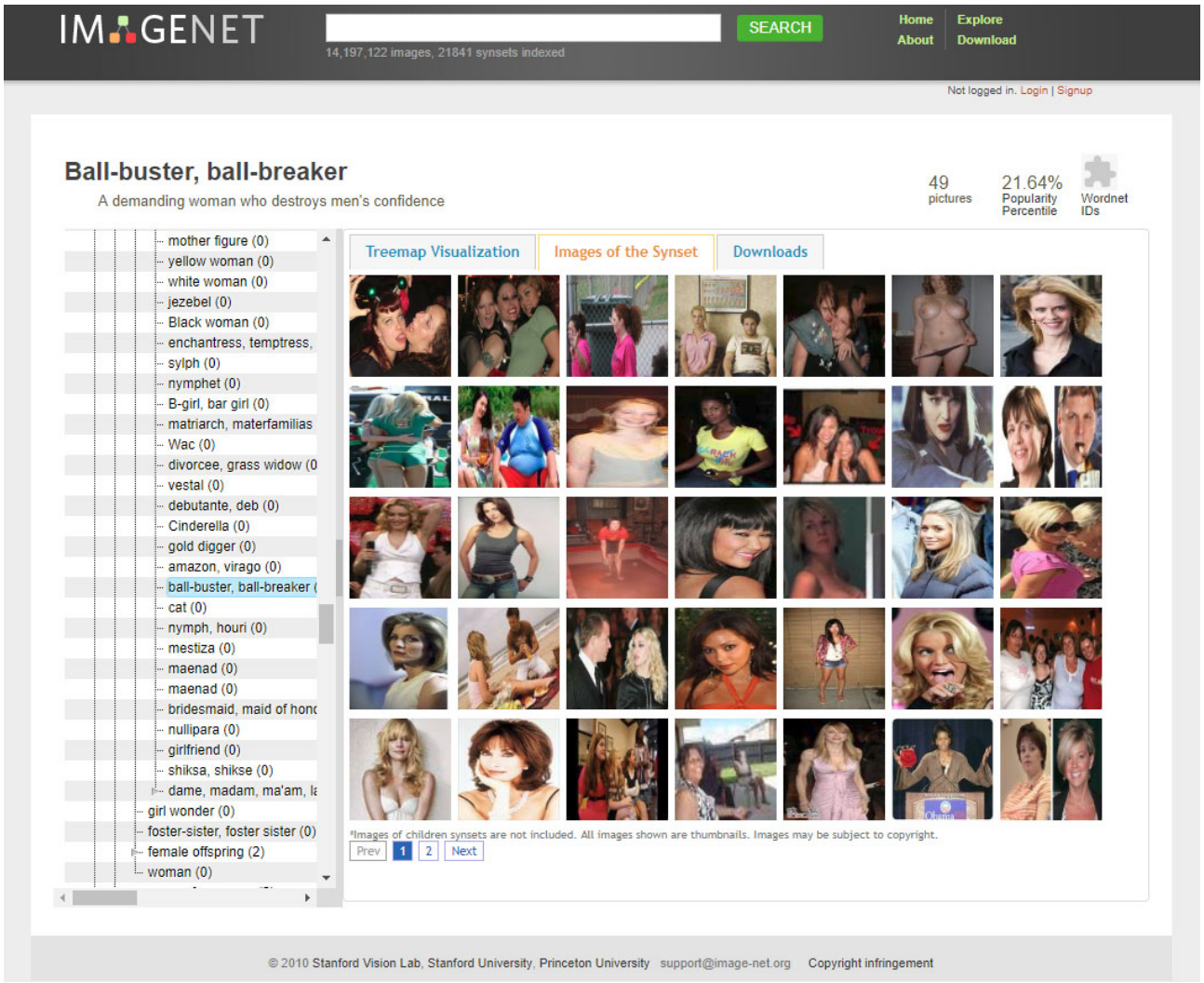
things is to divide an almost infinitely complex universe into separate phenomena. To impose order onto an undifferentiated mass, to ascribe phenomena to a category - that is, to name a thing - is in turn a means of reifying the existence of that category.

In the case of ImageNet, noun categories such as 'apple' or 'apple butter' might seem reasonably uncontroversial, but not all nouns are created equal. To borrow an idea from linguist George Lakoff, the concept of an "apple" is more nouny than the concept of 'light', which in turn is more nouny than a concept such as 'health'.[17] Nouns occupy various places on an axis from the concrete to the abstract, and from the descriptive to the judgmental. These gradients have been erased in the logic of ImageNet. Everything is flattened out and pinned to a label, like taxidermy butterflies in a display case. The results can be problematic, illogical, and cruel, especially when it comes to labels applied to people.

ImageNet contains 2,833 subcategories under the top-level category 'Person'. 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 worldview. ImageNet classifies people into a huge range of types including race, nationality, profession, economic status, behaviour, character and even morality. There are categories for racial and national identities including Alaska Native, Anglo-American, Black, Black African, Black Woman, Central American, Eurasian, German

American, Japanese, Lapp, Latin American, Mexican-American, Nicaraguan, Nigerian, Pakistani, Papuan, South American Indian, Spanish American, Texan, Uzbek, White, Yemeni and Zulu. Other people are labelled by their careers or hobbies: there are Boy Scouts, cheerleaders, cognitive neuroscientists, hairdressers, intelligence analysts, mythologists, retailers, retirees and so on.

As we go further into the depths of ImageNet's Person categories, the classifications of humans within it take a sharp and dark turn. There are categories for Bad Person, Call Girl, Drug Addict, Closet Queen, Convict, Crazy, Failure, Flop, Fucker, Hypocrite, Jezebel, Kleptomaniac, Loser, Melancholic, Nonperson, Pervert, Prima Donna, Schizophrenic, Second- Rater, Spinster, Streetwalker, Stud, Tosser, Unskilled Person, Wanton, Waverer and Wimp. There are many racist slurs and misogynistic terms.



Selections from the "Person" classes, ImageNet

Of course, ImageNet was typically used for object recognition - so the Person category was rarely discussed at technical conferences, nor has it received much public attention. However, this complex architecture of images of real people, tagged with often offensive labels, has been publicly available on the internet for a decade. It provides a powerful and important example of the complexities and dangers of human classification, and the sliding spectrum from supposedly unproblematic labels like 'trumpeter' or 'tennis player' to concepts like 'spastic', 'mulatto', or 'redneck'. Regardless of the supposed neutrality of any particular category, the selection of images skews the meaning in ways that are gendered, racialised, ableist and ageist. ImageNet is an object lesson, if you will, in what happens when people are categorised like objects. And this practice has only become more common in recent years, often inside the big AI companies, where there is no way for outsiders to see how images are being ordered and classified.

Finally, there is the issue of where the thousands of images in ImageNet's Person class were drawn from. By harvesting images en masse from image search engines like Google, ImageNet's creators appropriated people's selfies and vacation photos without their knowledge, and then labelled and repackaged them as the underlying data for much of an entire field.[18] When we take a look at the bedrock layer of labeled images, we find highly questionable semiotic assumptions, echoes of nineteenth- century phrenology, and the representational harm of classifying images of people without their consent or participation.

ImageNet Roulette: An Experiment in Classification

The ImageNet dataset is typically used for object recognition. But as part of our archeological method, we were interested to see what would happen if we trained an AI model exclusively on its 'person' categories. The result of that experiment is ImageNet Roulette.

ImageNet Roulette uses an open-source Caffe deep- learning framework (produced at UC Berkeley) trained on the images and labels in the 'person' categories (which are currently 'down for maintenance'). Proper nouns were removed.

When a user uploads a picture, the application first runs a face detector to locate any faces. If it finds any, it sends them to the Caffe model for classification. The application then returns the original images with a bounding box showing the detected face and the label the classifier has assigned to the image. If no faces are detected, the application sends the entire scene to the Caffe model and returns an image with a label in the upper left corner.

As we have shown, ImageNet contains a number of problematic, offensive and bizarre categories. Hence, the results ImageNet Roulette returns often draw upon those categories. That is by design: we want to shed light on what happens when technical systems are trained using problematic training data. AI classifications of people are rarely made visible to the people being classified. ImageNet Roulette provides a glimpse into that process - and to show how things can go wrong.

ImageNet Roulette does not store the photos people upload.

https://imagenet- roulette.paglen.com/

Labelled Images

Images are laden with potential meanings, irresolvable questions and contradictions. In trying to resolve these ambiguities, ImageNet's labels often compress and simplify images into deadpan banalities. One photograph shows a dark-skinned toddler wearing tattered and dirty clothes and clutching a soot-stained doll. The child's mouth is open. The image is completely devoid of context. Who is this child? Where is it? The photograph is simply labeled 'toy'.

But some labels are just nonsensical. A woman sleeps in an airplane seat, her right arm protectively curled around her pregnant stomach. The image is labeled 'snob'. A photoshopped picture shows a smiling Barack Obama wearing a Nazi uniform, his arm raised and holding a Nazi flag. It is labeled 'Bolshevik'.

At the image layer of the training set, like everywhere else, we find assumptions, politics and worldviews. According to ImageNet, for example, Sigourney Weaver is a 'hermaphrodite', a young man wearing a straw hat is a 'tosser', and a young woman lying on a beach towel is a 'kleptomaniac'. But the worldview of ImageNet isn't limited to the bizarre or derogatory conjoining of pictures and labels.

Other assumptions about the relationship between pictures and concepts recall physiognomy, the pseudoscientific assumption that something about a person's essential character can be gleaned by observing features of their body and face. ImageNet takes this to an extreme, assuming that whether someone is a 'debtor', a 'snob', a 'swinger', or a 'slav' can be determined by inspecting their photograph. In the weird metaphysics of ImageNet, there are separate image categories for 'assistant professor' and 'associate professor' - as though if someone were to get a promotion, their biometric signature would reflect the change in rank.

Of course, these sorts of assumptions have their own dark histories and attendant politics.

UTK: Making Race and Gender from Your Face

In 1839, the mathematician Franc?ois Arago claimed that through photographs, 'objects preserve mathematically their forms'.[19] Placed into the nineteenth-century context of imperialism and social Darwinism, photography helped to animate - and lend a 'scientific' veneer to - various forms of phrenology, physiognomy, and eugenics.[20] Physiognomists such as Francis Galton and Cesare Lombroso created composite images of criminals, studied the feet of prostitutes, measured skulls and compiled meticulous archives of labelled images and measurements, all in an effort to use 'mechanical'

processes to detect visual signals in classifications of race, criminality and deviance from bourgeois ideals. This was done to capture and pathologise what was seen as deviant or criminal behaviour, and make such behaviour observable in the world.

And as we shall see, not only have the underlying assumptions of physiognomy made a comeback with contemporary training sets, but a number of training sets are designed to use algorithms and facial landmarks as latter-day calipers to conduct contemporary versions of craniometry.

For example, the UTKFace dataset (produced by a group at the University of Tennessee at Knoxville) consists of over 20,000 images of faces with annotations for age, gender and race. The dataset's authors state that the dataset can be used for a variety of tasks, like automated face detection, age estimation and age progression.[21]



UTKFace Dataset

The annotations for each image include an estimated age for each person, expressed in years from zero to 116. Gender is a binary choice: either zero for male or one for female. Second, race is categorised from zero to four, and places people in one of five classes: White, Black, Asian, Indian, or 'Others'.

The politics here are as obvious as they are troubling. At the category level, the researchers' conception of gender is as a simple binary structure, with 'male' and 'female' the only alternatives. At the level of the image label is the assumption that someone's gender identity can be ascertained through a photograph.

Labels

The labels of each face image is embedded in the file name, formated like
`[age]_[gender]_[race]_[date&time].jpg`

- `[age]` is an integer from 0 to 116, indicating the age
- `[gender]` is either 0 (male) or 1 (female)
- `[race]` is an integer from 0 to 4, denoting White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern).
- `[date&time]` is in the format of yyyymmddHHMMSSFFF, showing the date and time an image was collected to UTKFace

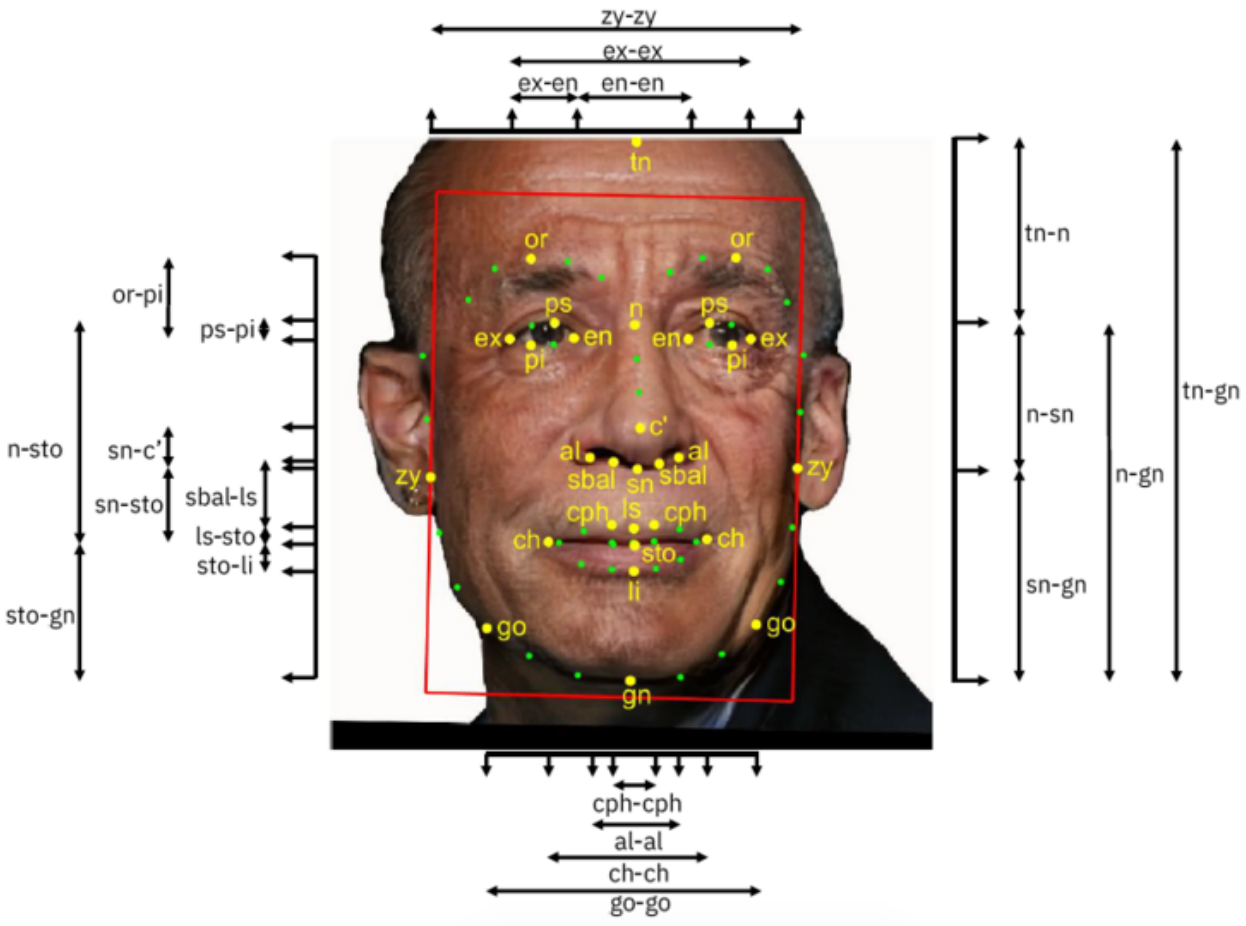
The classificatory schema for race recalls many of the deeply problematic racial classifications of the twentieth century. For example, the South African apartheid regime sought to classify the entire population into four categories: Black, White, Coloured, or Indian.[22] Around 1970, the South African government created a unified 'identity passbook' called The Book of Life, which linked to a centrally managed database created by IBM. These classifications were based on dubious and shifting criteria of 'appearance and general acceptance or repute', and many people were reclassified, sometimes multiple times. [23] The South African system of racial classification was intentionally very different from the American 'one-drop' rule, which stated that even one ancestor of African descent made somebody Black, likely because nearly all white South Africans had some traceable black African ancestry.[24] Above all, these systems of classifications caused enormous harm to people, and the elusive classifier of a pure 'race' signifier was always in dispute. However, seeking to improve matters by producing 'more diverse' AI training sets presents its own complications.

IBM'S Diversity in Faces

IBM's 'Diversity in Faces' dataset was created as a response to critics who had shown that the company's facial-recognition software often simply did not recognise the faces of people with darker skin.[25] IBM publicly promised to improve their facial-recognition datasets to make them more 'representative' and published the 'Diversity in Faces' (DiF) dataset as a result.[26] Constructed to be 'a computationally practical basis for ensuring fairness and accuracy in face recognition', the DiF consists of almost a million images of people pulled from the Yahoo! Flickr Creative Commons dataset, assembled specifically to achieve statistical parity among categories of skin tone, facial structure, age and gender.[27]

The dataset itself continued the practice of collecting hundreds of thousands of images of unsuspecting people who had uploaded pictures to sites like Flickr.[28] But the dataset contains a unique set of categories not previously seen in other face-image datasets. The IBM DiF team asks whether age, gender and skin colour are truly sufficient in generating a dataset that can ensure fairness and accuracy and concludes that even more classifications are needed. So they move into truly strange territory: including facial symmetry and skull shapes to build a complete picture of the face. The researchers claim that the use of craniofacial features is justified because it captures much more granular information about a person's face than just gender, age and skin colour alone. The paper accompanying the dataset specifically highlights prior work done to show that skin colour is itself a weak predictor of race, but this begs the question of why moving to skull shapes is appropriate.

Craniometry was a leading methodological approach of biological determinism during the nineteenth century. As Stephen Jay Gould shows in his book *The Mismeasure of Man*, skull size was used by nineteenth- and twentieth-century pseudoscientists as a spurious way to claim inherent superiority of white people over black people, and different skull shapes and weights were said to determine people's intelligence - always along racial lines.[29]



IBM's Diversity in Faces

While the efforts of companies to build more diverse training sets is often put in the language of increasing 'fairness' and 'mitigating bias', clearly there are strong business imperatives to produce tools that will work more effectively across wider markets. However, here too the technical process of categorising and classifying people is shown to be a political act. For example, how is a 'fair' distribution achieved within the dataset?

IBM decided to use a mathematical approach to quantifying 'diversity' and 'evenness', so that a consistent measure of evenness exists throughout the dataset for every feature quantified. The dataset also contains subjective annotations for age and gender, which are generated using three independent Amazon Turk workers for each image, similar to the methods used by ImageNet.[30] So people's gender and age are being 'predicted' based on three clickworkers' guesses about what's shown in a photograph scraped from the internet. It harkens back to the early carnival game of 'Guess Your Weight!', with similar levels of scientific validity.

Ultimately, beyond these deep methodological concerns, the concept and political history of diversity is being drained of its meaning and left to refer merely to expanded biological

phenotyping. Diversity in this context just means a wider range of skull shapes and facial symmetries. For computer vision researchers, this may seem like a 'mathematization of fairness', but it simply serves to improve the efficiency of surveillance systems. And even after all these attempts at expanding the ways which people are classified, the Diversity in Faces set still relies on a binary classification for gender: people can only be labelled male or female. Achieving parity amongst different categories is not the same as achieving diversity or fairness, and IBM's data construction and analysis

perpetuates a harmful set of classifications within a narrow worldview.

Epistemics of Training Sets

What are the assumptions undergirding visual AI systems? First, the underlying theoretical paradigm of the training sets assumes that concepts - whether 'corn', 'gender', 'emotions' or 'losers' - exist in the first place, and that those concepts are fixed, universal, and have some sort of transcendental grounding and internal consistency. Second, it assumes a fixed and universal correspondence between images and concepts, appearances and essences. What's more, it assumes uncomplicated, self- evident and measurable ties between images, referents and labels. In other words, it assumes that different concepts - whether 'corn' or 'kleptomaniacs' - have some kind of essence that unites each instance of them, and that that underlying essence expresses itself visually. Moreover, the theory goes, that visual essence is discernible by using statistical methods to look for formal patterns across a collection of labeled images. Images of people dubbed 'losers', the theory goes, contain some kind of visual pattern that distinguishes them from, say, 'farmers', 'assistant professors', or, for that matter, apples. Finally, this approach assumes that all concrete nouns are created equally, and that many abstract nouns also express themselves concretely and visually (i.e., 'happiness' or 'anti- Semitism).

The training sets of labelled images that are ubiquitous in contemporary computer vision and AI are built on a foundation of unsubstantiated and unstable epistemological and metaphysical assumptions about the nature of images, labels, categorisation and representation. Furthermore, those epistemological and metaphysical assumptions hark back to historical approaches where people were visually assessed and classified as a tool of oppression and race science.

Datasets aren't simply raw materials to feed algorithms, but are political interventions. As such, much of the discussion around 'bias' in AI systems misses the mark: there is no 'neutral', 'natural', or 'apolitical' vantage point that training data can be built upon. There is no easy technical 'fix' by shifting demographics, deleting offensive terms, or seeking equal representation by skin tone. The whole endeavour of collecting images, categorising them, and labelling them is itself a form of politics, filled with questions about who gets to decide what images mean and what kinds of social and political work those representations perform.

Missing Persons

In January 2019, images in ImageNet's 'Person' category began disappearing. Suddenly, 1.2 million photos were no longer accessible on Stanford University's servers. Gone were the pictures of cheerleaders, scuba divers, welders, altar boys, retirees and pilots. The picture of a man drinking beer characterised as an 'alcoholic' disappeared, as did the pictures of a woman in a bikini dubbed a 'slattern' and a young boy classified as a 'loser'. The picture of a man eating a sandwich (labelled a 'selfish person') met the same fate. When you search for these images, the ImageNet website responds with a statement that it is under maintenance, and only the categories used in the ImageNet competition are still included in the search results.

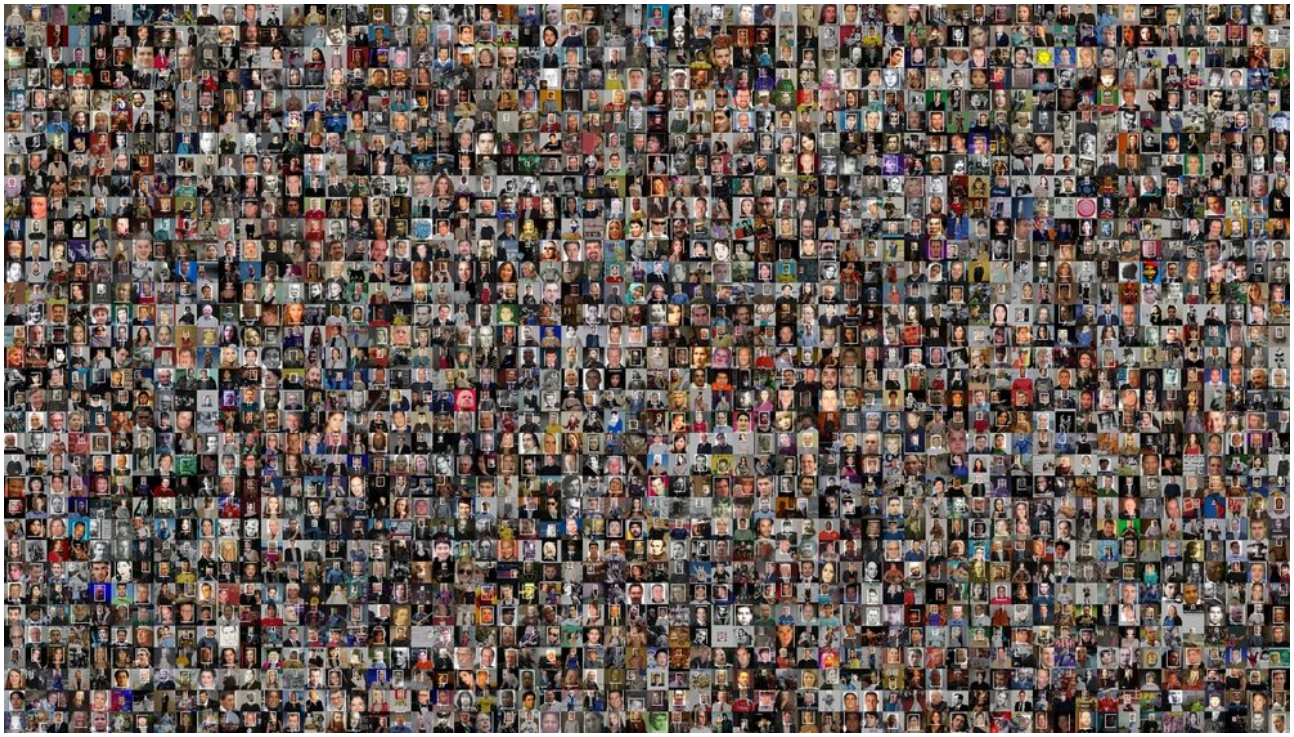
But once it came back online, the search functionality on the site was modified so that it would only return results for categories that had been included in ImageNet's annual computer-vision contest. As of this writing, the 'Person' category is still browsable from the data set's online interface, but the images fail to load. The URLs for the original images are still downloadable.[31]

Over the next few months, other image collections used in computer-vision and AI research also began to disappear. In response to research published by Adam Harvey and Jules LaPlace,[32] Duke University took down a massive photo repository of surveillance-camera footage of students attending classes (called the Duke Multi-Target, Multi-Camera [MTMC] dataset). It turned out that the authors of the dataset had violated the terms of their Institutional Review Board approval by collecting images from people in public space, and by making their dataset publicly available. [33]

Similar datasets created from surveillance footage disappeared from servers at the University of

Colorado Colorado Springs, and more from Stanford University, where a collection of faces culled from a webcam installed at San Francisco's iconic Brainwash Cafe was 'removed from access at the request of the depositor'.[34]

By early June, Microsoft had followed suit, removing their landmark "MS-CELEB" collection of approximately ten million photos from 100,000 people scraped from the internet in 2016. It was the largest public facial- recognition dataset in the world, and the people included were not just famous actors and politicians, but also journalists, activists, policy makers, academics, and artists.[35] Ironically, several of the people who had been included in the set without any consent are known for their work critiquing surveillance and facial recognition itself, including filmmaker Laura Poitras, digital rights activist Jillian York, critic Evgeny Morozov and author of *Surveillance Capitalism* Shoshana Zuboff. After an investigation in the *Financial Times* based on Harvey and LaPlace's work was published, the set disappeared. [36] A spokesperson for Microsoft claimed simply that it was removed 'because the research challenge is over'.[37]



MS CELEB dataset

On one hand, removing these problematic datasets from the internet may seem like a victory. The most obvious privacy and ethical violations are addressed by making them no longer accessible. However, taking them offline doesn't stop their work in the world: these training sets have been downloaded countless times, and have made their way into many production AI systems and academic papers. By erasing them completely, not only is a significant part of the history of AI lost, but researchers are unable to see how the assumptions, labels and classificatory approaches have been replicated in new systems, or trace the provenance of skews and biases exhibited in working systems. Facial-recognition and emotion-recognition AI systems are already propagating into hiring, education and healthcare. They are part of security checks at airports and interview protocols at Fortune 500 companies. Not being able to see the basis on which AI systems are trained removes an important forensic method to understand how they work. This has serious consequences.

For example, a recent paper led by a PhD student at the University of Cambridge introduced a real-time drone surveillance system to identify violent individuals in public areas. It is trained on datasets of

'violent behaviour' and uses those models for drone surveillance systems to detect and isolate violent behaviour in crowds. The team created the Aerial Violent Individual (AVI) Dataset, which consists of 2,000 images of people engaged in five activities: punching, stabbing, shooting, kicking and strangling. In order to train their AI, they asked twenty-five volunteers between the ages of eighteen and twenty-five to mimic these actions. Watching the videos is almost comic. The actors stand far apart and perform strangely exaggerated gestures. It looks like a children's pantomime, or badly modelled game characters.[38] The full dataset is not available for the public to download. The lead researcher, Amarjot Singh (now at Stanford University), said he plans to test the AI system by flying drones over two major festivals, and potentially at national borders in India. [39] [40]

An archeological analysis of the AVI dataset - similar to our analyses of ImageNet, JAFFE, and Diversity in Faces - could be very revealing. There is clearly a significant difference between staged performances of violence and real-world cases. The researchers are training drones to recognise pantomimes of violence, with all of the misunderstandings that might come with that. Furthermore, the AVI dataset doesn't have anything for 'actions that aren't violence but might look like it'; neither do they publish any details about their false-positive rate (how often their system detects nonviolent behavior as violent).[41] Until their data is released, it is impossible to do forensic testing on how they classify and interpret human bodies, actions or inactions.

This is the problem of inaccessible or disappearing datasets. If they are, or were, being used in systems that play a role in everyday life, it is important to be able to study and understand the worldview they normalise. Developing frameworks within which future researchers can access these data sets in ways that don't perpetuate harm is a topic for further work.

Conclusion: Who decides?

The Lombrosian criminologists and other phrenologists of the early twentieth century didn't see themselves as political reactionaries. On the contrary, as Steven Jay Gould points out, they tended to be liberals and socialists whose intention was 'to use modern science as a cleansing broom to sweep away from jurisprudence the outdated philosophical baggage of free will and unmitigated moral responsibility'.[42] They believed their anthropometric method of studying criminality could lead to a more enlightened approach to the application of justice. Some of them truly believed they were 'de-biasing' criminal justice systems, creating 'fairer' outcomes through the application of their 'scientific' and 'objective' methods.

Amid the heyday of phrenology and 'criminal anthropology', the artist Rene? Magritte completed a painting of a pipe and coupled it with the words 'Ceci n'est pas une pipe'. Magritte called the painting *La trahison des images*, 'The Treachery of Images'. That same year, he penned a text in the surrealist newsletter *La Re?volution surre?aliste*. 'Les mots et les images' is a playful romp through the complexities and subtleties of images, labels, icons and references, underscoring the extent to which there is nothing at all straightforward about the relationship between images and words or linguistic concepts. The series would culminate in a series of paintings: 'This Is Not an Apple'.

The contrast between Magritte and the physiognomists' approach to representation speaks to two very different conceptions of the fundamental relationship between images and their labels, and of representation itself. For the physiognomists, there was an underlying faith that the relationship between an image of a person and the character of that person was inscribed in the images themselves. Magritte's assumption was almost diametrically opposed: that images in and of themselves have, at best, a very unstable relationship to the things they seem to represent, one that can be sculpted by whoever has the power to say what a particular image means.

For Magritte, the meaning of images is relational, open to contestation. At first blush, his painting might seem like a simple semiotic stunt, but the underlying dynamic Magritte underlines in the painting points to a much broader politics of representation and self-representation.



Memphis Sanitation Workers Strike of 1968

Struggles for justice have always been, in part, struggles over the meaning of images and representations. In 1968, African American sanitation workers went on strike to protest dangerous working conditions and terrible treatment at the hands of Memphis's racist government. They held up signs recalling language from the nineteenth-century abolitionist movement: 'I AM A MAN'. In the 1970s, queer-liberation activists appropriated a symbol originally used in Nazi concentration camps to identify prisoners who had been labeled as homosexual, bisexual, and transgender. The pink triangle became a badge of pride, one of the most iconic symbols of queer-liberation movements. Examples such as these - of people trying to define the meaning of their own representations - are everywhere in struggles for justice. Representations aren't simply confined to the spheres of language and culture, but have real implications in terms of rights, liberties, and forms of self-determination.

There is much at stake in the architecture and contents of the training sets used in AI. They can promote or discriminate, approve or reject, render visible or invisible, judge or enforce. And so we need to examine them - because they are already used to examine us - and to have a wider public discussion about their consequences, rather than keeping it within academic corridors. As training sets are increasingly part of our urban, legal, logistical, and commercial infrastructures they have an important but under examined role: the power to shape the world in their own images.

Credits and Acknowledgements

Authors: Kate Crawford and Trevor Paglen

Published by The AI Now Institute, NYU (https://ainowinstitute.org /)

Full citation: Kate Crawford and Trevor Paglen, "Excavating AI: The Politics of Training Sets for Machine Learning (September 19, 2019) https://excavating.ai

Acknowledgements: thanks to all those who have given editorial feedback, technical support, research contributions and conversations on these issues over the years, including Arvind Narayanan, Daniel Neves, Varoon Mathur, Olga Russakovsky, Leif Ryge, Lea Saint- Raymond and Kiran Samuel.

Additional thanks to Mario Mainetti and Carlo Barbatti and all the staff at the Fondazione Prada, and to Alona Pardo and the staff at the Barbican Centre. The images in this essay and many more are part of the Fondazione Prada Osservatorio *Training Humans* exhibition, in Milan from 12 September, 2019 through 24 February 2020; and at the Barbican Centre in London as part of the exhibition *From Apple to Anomaly (Pictures and Labels)* from 26 September, 2019 through 16 February, 2020.

Credit for ImageNet Roulette software: developed by Leif Ryge at Paglen Studio.

'Excavating AI: The Politics of Training Sets for Machine Learning' (September 19, 2019) https://excavating.ai by Kate Crawford and Trevor Paglen was originally published by The AI Now Institute, NYU (https://ainowinstitute.org/) and is republished with kind permission of Kate Crawford and Trevor Paglen.

---

[1] Minsky currently faces serious allegations related to convicted paedophile and rapist Jeffrey Epstein. Minsky was one of several scientists who met with Epstein and visited his island retreat where underage girls were forced to have sex with members of Epstein's coterie. As scholar Meredith Broussard observed, this was part of a broader culture of exclusion that became endemic in AI: 'as wonderfully creative as Minsky and his cohort were, they also solidified the culture of tech as a billionaire boys' club. Math, physics, and the other "hard" sciences have never been hospitable to women and people of color; tech followed this lead.' See Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World* (Cambridge, Massachusetts and London: MIT Press, 2018), p. 174.

[2] See Daniel Crevier, *AI: The Tumultuous History of the Search for Artificial Intelligence* (New York: Basic Books, 1993), p. 88.

[3] Minsky gets the credit for this idea, but clearly Papert, Sussman, and teams of 'summer workers' were all part of this early effort to get computers to describe objects in the world. See Seymour A. Papert, 'The Summer Vision Project' (July 1, 1966), https://dspace.mit.edu/handle/1721.1/6125. As he wrote: 'The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition"'.

[4] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed, Prentice Hall Series in Artificial Intelligence (Upper Saddle River, NJ: Prentice Hall, 2010), p. 987.

[5] In the late 1970s, Ryszard Michalski wrote an algorithm based on 'symbolic variables' and logical rules. This language was very popular in the 1980s and 1990s, but, as the rules of decision-making and qualification became more complex, the language became less usable. At the same moment, the potential of using large training sets triggered a shift from this conceptual clustering to contemporary machine-learning approaches. See Ryszard Michalski, 'Pattern Recognition as Rule-Guided Inductive Inference'. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2 (1980), pp. 349–361.

[6] There are hundreds of scholarly books in this category, but for a good place to start, see W.J.T. Mitchell, *Picture Theory: Essays on Verbal and Visual Representation*, Paperback ed., [Nachdr.] (Chicago: University of Chicago Press, 2007).

[7] M. Lyons et al., 'Coding Facial Expressions with Gabor Wavelets', in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* (Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan: IEEE Comput. Soc, 1998), pp. 200– 205, https://doi.org/10.1109/AFGR.1998.670949.

[8] As described in the 'AI Now Report' (2018), this classification of emotions into six categories has its root in the work of the psychologist Paul Ekman. 'Studying faces, according to Ekman, produces an objective reading of authentic interior states - a direct window to the soul. Underlying his belief was the idea that emotions are fixed and universal, identical across individuals, and clearly visible in observable biological mechanisms regardless of cultural context. But Ekman's work has been deeply criticized by psychologists, anthropologists, and other researchers who have found his theories do not hold up under sustained scrutiny. The psychologist Lisa Feldman Barrett and her colleagues have argued that an understanding of emotions in terms of these rigid categories and simplistic physiological causes is no longer tenable. Nonetheless, AI researchers have taken his work as fact,

and used it as a basis for automating emotion detection.' Meredith Whitaker et al., 'AI Now Report 2018', AI Now Institute (December 2018), https://ainowinstitute.org/AI_Now_2018_Report.pdf. See also Lisa Feldman Barrett et al., 'Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements', *Psychological Science in the Public Interest* 20 (1) (July 17, 2019), pp. 1–68, https://doi.org/10.1177/1529100619832930.

[9] See, for example, Ruth Leys, 'How Did Fear Become a Scientific Object and What Kind of Object Is It?', *Representations* 110 (1) (May 2010), pp. 66–104, https://doi.org/10.1525/rep.2010.110.1.66. Leys has offered a number of critiques of Ekman's research programme, most recently in Ruth Leys, *The Ascent of Affect: Genealogy and Critique* (Chicago and London: University of Chicago Press, 2017). See also Lisa Feldman Barrett, 'Are Emotions Natural Kinds?', *Perspectives on Psychological Science* 1 (1) (March 2006), pp. 28–58, https://doi.org/10.1111/j.1745- 6916.2006.00003.x; Erika H. Siegel et al., 'Emotion Fingerprints or Emotion Populations? A Meta-Analytic Investigation of Autonomic Features of Emotion Categories.', *Psychological Bulletin,* 20180201, https://doi.org/10.1037/bul0000128.

[10] Fei-Fei Li, as quoted in Dave Gershgorn, 'The Data That Transformed AI Research - and Possibly the World', *Quartz* (July 26, 2017), https://qz.com/1034972/the- data-that-changed-the-direction-of-ai-research-and-possibly-the-world/. Emphasis added.

[11] John Markoff, 'Seeking a Better Way to Find Web Images', *The New York Times (*November, 19, 2012), sec. Science, https://www.nytimes.com/2012/11/20/s cience/for-web-images-creat-ng-new-technology-to-seek-and-find.html.

[12] Their paper can be found here: Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, 'ImageNet Classification with Deep Convolutional Neural Networks', in *Advances in Neural Information Processing Systems* 25, ed. F. Pereira et al. (Curran Associates, Inc., 2012), pp. 1097–1105, http://papers.nips.cc/paper/4824-imagenet-classification-with-deep- convolutional-neura--networks.pdf.

[13] Released in the mid-1980s, this lexical database for the English language can be seen as a thesaurus that defines and groups English words into synsets, i.e., sets of synonyms. https://wordnet.princeton.edu. This project takes place in a broader history of computational linguistics and natural-language processing (NLP), which developed during the same period. This subfield aims at programming computers to process and analyse large amounts of natural language data, using machine-learning algorithms.

[14] See Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences*, First paperback edition, Inside Technology (Cambridge, Massachusetts and London: MIT Press, 2000), pp. 44, 107; Anja Bechmann and Geoffrey C. Bowker, 'Unsupervised by Any Other Name: Hidden Layers of Knowledge Production in Artificial Intelligence on Social Media', *Big Data & Society* 6 (1) (January 2019): 205395171881956, https://doi.org/10.1177/2053951718819569.

[15] These are some of the categories that have now been entirely deleted from ImageNet as of January, 24, 2019.

[16] For an account of the politics of classification in the Library of Congress, see Sanford Berman, *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People* (Metuchen, NJ: Scarecrow Press, 1971).

[17] We're drawing in part here on the work of George Lakoff in *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* (Chicago: University of Chicago Press, 2012).

[18] See Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, 'Imagenet: A Large-Scale Hierarchical Image Database' In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–55.

[19] Quoted in Allan Sekula, 'The Body and the Archive', *October* 39 (1986), pp. 3–64, https://doi.org/10.2307/778312.

[20] Ibid; for a broader discussion of objectivity, scientific judgment, and a more nuanced take on photography's role in it, see Lorraine Daston and Peter Galison, *Objectivity,* Paperback ed. (New York: Zone Books, 2010).

[21] 'UTKFace – Aicip', accessed August 28, 2019, http://aicip.eecs.utk.edu/wiki/UTKFace.

[22] See Paul N. Edwards and Gabrielle Hecht, 'History and the Technopolitics of Identity: The Case of Apartheid South Africa', *Journal of Southern African Studies* 36 (3) (September 2010), pp. 619–39, https://doi.org/10.1080/03057070.2010.507568. Earlier classifications used in the 1950

Population Act and Group Areas Act used four classes: 'Europeans, Asiatics, persons of mixed race or coloureds, and "natives" or pure-blooded individuals of the Bantu race' (Bowker and Star, p. 197). Black South Africans were required to carry pass books and could not, for example, spend more than 72 hours in a white area without permission from the government for a work contract (p. 198).

[23] Bowker and Star, 208.

[24] See F. James Davis, *Who Is Black? One Nation's Definition*, 10th anniversary ed. (University Park, PA: Pennsylvania State University Press, 2001).

[25] See Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', in Conference on Fairness, Accountability, and Transparency (2018), pp. 77–91, http://proceedings.mlr.press/v81/buolamwini18a.html.

[26] Michele Merler et al., 'Diversity in Faces', ArXiv:1901.10436 [Cs] (January 29, 2019), http://arxiv.org/abs/1901.10436.

[27] 'Webscope | Yahoo Labs', accessed August 28, 2019, https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67&guccounter=1.

[28] Olivia Solon, 'Facial Recognition's "Dirty Little Secret": Millions of Online Photos Scraped without Consent' (March 12, 2019), https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921.

[29] Stephen Jay Gould, *The Mismeasure of Man*, revised and expanded (New York: Norton, 1996). The approach of measuring intelligence based on skull size was prevalent across Europe and the US. For example, in France, Paul Broca and Gustave Le Bon developed the approach of measuring intelligence based on skull size. See Paul Broca, 'Sur le crane de Schiller et sur l'indice cubique des cranes', *Bulletin de la Societe? d'anthropologie de Paris*, I° Serie, t. 5, fasc. 1, pp. 253-60, 1864. Gustave Le Bon, *L'homme et les societes. Leurs origines et leur de?veloppement* (Paris: Edition J. Rothschild, 1881). In Nazi Germany, the 'anthropologist' Eva Justin wrote about Sinti and Roma people, based on anthropometric and skull measurements. See Eva Justin, Lebensschicksale artfremd erzogener Zigeunerkinder und ihrer Nachkommen [Biographical destinies of Gypsy children and their offspring who were educated in a manner inappropriate for their species], doctoral dissertation, Friedrich-Wilhelms-Universitat Berlin, 1943.

[30] 'Figure Eight | The Essential High- Quality Data Annotation Platform', Figure Eight, accessed August 28, 2019, https://www.figure-eight.com/.

[31] The authors made a backup of the ImageNet dataset prior to much of its deletion.

[32] Their 'MegaPixels' project is here: https://megapixels.cc/.

[33] Jake Satisky, 'A Duke Study Recorded Thousands of Students' Faces. Now They're Being Used All over the World', *The Chronicle* (June 12, 2019), https://www.dukechronicle.com/article/2019/06/duke-university-facial-recognition-data-set-study-surveillance-video-students-china-uyghur.

[34] '2nd Unconstrained Face Detection and Open Set Recognition Challenge', accessed August 28, 2019, https://vast.uccs.edu/Opensetface/; Russell Stewart, Brainwash Dataset (Stanford Digital Repository, 2015), https://purl.stanford.edu/sx925dc9385.

[35] Melissa Locker, 'Microsoft, Duke, and Stanford Quietly Delete Databases with Millions of Faces', Fast Company (June 6, 2019), https://www.fastcompany.com/90360490/ms-celeb-microsoft-deletes-10m-faces-from-face-database.

[36] Madhumita Murgia, 'Who's Using Your Face? The Ugly Truth about Facial Recognition', *Financial Times* (April 19, 2019), https://www.ft.com/content/cf19b956-60a2-11e9-b-85-3acd5d43599e.

[37] Locker, 'Microsoft, Duke, and Stanford Quietly Delete Databases'.

[38] Full video here: Amarjot Singh, 'Eye in the Sky: Real-Time Drone Surveillance System (DSS) for Violent Individuals Identification' (2018), https://www.youtube.com/watch?time_continue=1&v=zYypJPJipYc.

[39] Steven Melendez, 'Watch This Drone Use AI to Spot Violence in Crowds from the Sky', *Fast Company* (June 6, 2018), https://www.fastcompany.com/40581669/watch-this-drone-use-ai-t-

-spot-violence-from-the-sky.

[40] James Vincent, 'Drones Taught to Spot Violent Behavior in Crowds Using AI', *The Verge* (June 6, 2018), https://www.theverge.com/2018/6/6/17433482/ai-auto...drones-spot-violent-behavior-crowds.

[41] Ibid.

[42] Gould, *The Mismeasure of Man*, p. 140.

Kate Crawford and Trevor Paglen

Kate Crawford is a leading academic focusing on the social and political implications of artificial intelligence. For over a decade, her work has centred on understanding large-scale data systems in the wider contexts of politics, history, labour, and the environment. Kate Crawford is a research professor at USC Annenberg, and the Visiting Chair of AI and Justice at the École Normale Supérieure. In 2020, she is the inaugural Visiting Chair for AI and Justice at the École Normale Supérieure in Paris. She co-founded the AI Now Institute at New York University, a university centre dedicated to researching the social implications of AI and related technologies. In 2019, she and Vladan Joler won the Beazley Design of the Year Award, for their Anatomy of an AI System project, which was recently acquired by MoMA for its permanent collection. Crawford was also jointly awarded the Ayrton Prize from the British Society for the History of Science for the project Excavating AI. Her new book Atlas of AI is forthcoming with Yale University Press in 2021.

Trevor Paglen is an artist whose work spans image-making, sculpture, investigative journalism, writing, engineering and numerous other disciplines. Among his chief concerns are learning how to see the historical moment in which we live, and developing the means to imagine alternative futures.Paglen has had one-person exhibitions at Nam June Paik Art Center, Seoul; Museo Tamayo, Mexico City; the Nevada Museum of Art, Reno; Vienna Secession; Eli & Edythe Broad Art Museum, East Lansing, Michigan; Van Abbe Museum, Eindhoven; Frankfurter Kunstverein and Protocinema Istanbul, and participated in group exhibitions at the Metropolitan Museum of Art, New York;the San Francisco Museum of Modern Art; Tate Modern, London, and numerous other venues. He has launched an artwork into distant orbit around Earth in collaboration with Creative Time and MIT, contributed research and cinematography to the Academy Award-winning film Citizenfour, and created a radioactive public sculpture for the exclusion zone in Fukushima, Japan. He is the author of five books and numerous articles on subjects including experimental geography, state secrecy, military symbology, photography and visuality. Paglen's work has been profiled in the New York Times, Vice Magazine, the New Yorker, and Art Forum. In 2014, he received the Electronic Frontier Foundation's Pioneer Award for his work as a 'groundbreaking investigative artist'.Paglen holds a BA from UC Berkeley, an MFA from the Art Institute of Chicago, and a Ph.D. in Geography from UC Berkeley.