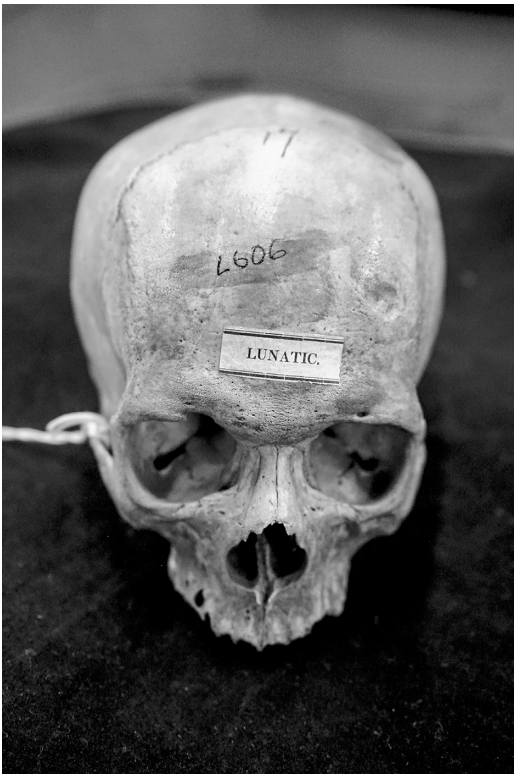
4
Classification

I am surrounded by human skulls. This room contains almost five hundred, collected in the early decades of the 1800s. All are varnished, with numbers inscribed in black ink on the frontal bone. Delicate calligraphic circles mark out areas of the skull associated in phrenology with particular qualities, including "Benevolence" and "Veneration." Some bear descriptions in capital letters, with words like "Dutchman," "Peruvian of the Inca Race," or "Lunatic." Each was painstakingly weighed, measured, and labeled by the American craniologist Samuel Morton. Morton was a physician, natural historian, and member of the Academy of Natural Sciences of Philadelphia. He gathered these human skulls from around the world by trading with a network of scientists and skull hunters who brought back specimens for his experiments, sometimes by robbing graves.[1] By the end of his life in 1851, Morton had amassed more than a thousand skulls, the largest collection in the world at the time.[2] Much of the archive is now held in storage at the Physical Anthropology Section of the Penn Museum in Philadelphia.

Morton was not a classical phrenologist in that he didn't believe that human character could be read through examin-



A skull from the Morton cranial collection marked "Lunatic." Photograph by Kate Crawford

ing the shape of the head. Rather, his aim was to classify and rank human races "objectively" by comparing the physical characteristics of skulls. He did this by dividing them into the five "races" of the world: African, Native American, Caucasian, Malay, and Mongolian—a typical taxonomy of the time and a reflection of the colonialist mentality that dominated its geopolitics.[3] This was the viewpoint of polygenism—the belief that distinct human races had evolved separately at different

times—legitimized by white European and American scholars and hailed by colonial explorers as a justification for racist violence and dispossession.[4] Craniometry grew to be one of their leading methods since it purported to assess human difference and merit accurately.[5]

Many of the skulls I see belong to people who were born in Africa but who died enslaved in the Americas. Morton measured these skulls by filling the cranial cavities with lead shot, then pouring the shot back into cylinders and gauging the volume of lead in cubic inches.[6] He published his results, comparing them to skulls he acquired from other locations: for example, he claimed that white people had the largest skulls, while Black people were on the bottom of the scale. Morton's tables of average skull volume by race were regarded as the cutting edge of science of the time. His work was cited for the rest of the century as objective, hard data that proved the relative intelligence of human races and biological superiority of the Caucasian race. This research was instrumented in the United States to maintain the legitimacy of slavery and racial segregation.[7] Considered the scientific state of the art at the time, it was used to authorize racial oppression long after the studies were no longer cited.

But Morton's work was not the kind of evidence it claimed to be. As Stephen Jay Gould describes in his landmark book *The Mismeasure of Man:*

> In short, and to put it bluntly, Morton's summaries are a patchwork of fudging and finagling in the clear interest of controlling *a priori* convictions. Yet—and this is the most intriguing aspect of his case—I find no evidence of conscious fraud. . . . The prevalence of unconscious finagling, on the other hand, suggests a general conclusion about the so-

> cial context of science. For if scientists can be honestly self-deluded to Morton's extent, then prior prejudice may be found anywhere, even in the basics of measuring bones and toting sums.[8]

Gould, and many others since, has reweighed the skulls and reexamined Morton's evidence.[9] Morton made errors and miscalculations, as well as procedural omissions, such as ignoring the basic fact that larger people have larger brains.[10] He selectively chose samples that supported his belief of white supremacy and deleted the subsamples that threw off his group averages. Contemporary assessments of the skulls at the Penn Museum show no significant differences among people—even when using Morton's data.[11] But prior prejudice—a way of seeing the world—had shaped what Morton believed was objective science and was a self-reinforcing loop that influenced his findings as much as the lead-filled skulls themselves.

Craniometry was, as Gould notes, "the leading numerical science of biological determinism during the nineteenth century" and was based on "egregious errors" in terms of the core underlying assumptions: that brain size equated to intelligence, that there are separate human races which are distinct biological species, and that those races could be placed in a hierarchy according to their intellect and innate character.[12] Ultimately, this kind of race science was debunked, but as Cornel West has argued, its dominant metaphors, logics, and categories not only supported white supremacy but also made specific political ideas about race possible while closing down others.[13]

Morton's legacy foreshadows epistemological problems with measurement and classification in artificial intelligence. Correlating cranial morphology with intelligence and claims to legal rights acts as a technical alibi for colonialism and slavery.[14] While there is a tendency to focus on the errors in skull mea-

surements and how to correct for them, the far greater error is in the underlying worldview that animated this methodology. The aim, then, should be not to call for more accurate or "fair" skull measurements to shore up racist models of intelligence but to condemn the approach altogether. The practices of classification that Morton used were *inherently* political, and his invalid assumptions about intelligence, race, and biology had far-ranging social and economic effects.

The politics of classification is a core practice in artificial intelligence. The practices of classification inform how machine intelligence is recognized and produced from university labs to the tech industry. As we saw in the previous chapter, artifacts in the world are turned into data through extraction, measurement, labeling, and ordering, and this becomes—intentionally or otherwise—a slippery ground truth for technical systems trained on that data. And when AI systems are shown to produce discriminatory results along the categories of race, class, gender, disability, or age, companies face considerable pressure to reform their tools or diversify their data. But the result is often a narrow response, usually an attempt to address technical errors and skewed data to make the AI system appear more fair. What is often missing is a more fundamental set of questions: How does classification function in machine learning? What is at stake when we classify? In what ways do classifications interact with the classified? And what unspoken social and political theories underlie and are supported by these classifications of the world?

In their landmark study of classification, Geoffrey Bowker and Susan Leigh Star write that "classifications are powerful technologies. Embedded in working infrastructures they become relatively invisible without losing any of their power."[15] Classification is an act of power, be it labeling images in AI training sets, tracking people with facial recognition, or pour-

ing lead shot into skulls. But classifications can disappear, as Bowker and Star observe, "into infrastructure, into habit, into the taken for granted."[16] We can easily forget that the classifications that are casually chosen to shape a technical system can play a dynamic role in shaping the social and material world.

The tendency to focus on the issue of bias in artificial intelligence has drawn us away from assessing the core practices of classification in AI, along with their attendant politics. To see that in action, in this chapter we'll explore some of the training datasets of the twenty-first century and observe how their schemas of social ordering naturalize hierarchies and magnify inequalities. We will also look at the limits of the bias debates in AI, where mathematical parity is frequently proposed to produce "fairer systems" instead of contending with underlying social, political, and economic structures. In short, we will consider how artificial intelligence uses classification to encode power.

Systems of Circular Logic

A decade ago, the suggestion that there could be a problem of bias in artificial intelligence was unorthodox. But now examples of discriminatory AI systems are legion, from gender bias in Apple's creditworthiness algorithms to racism in the COMPAS criminal risk assessment software and to age bias in Facebook's ad targeting.[17] Image recognition tools miscategorize Black faces, chatbots adopt racist and misogynistic language, voice recognition software fails to recognize female-sounding voices, and social media platforms show more highly paid job advertisements to men than to women.[18] As scholars like Ruha Benjamin and Safiya Noble have shown, there are hundreds of examples throughout the tech ecosystem.[19] Many more have never been detected or publicly admitted.

The typical structure of an episode in the ongoing AI bias narrative begins with an investigative journalist or whistleblower revealing how an AI system is producing discriminatory results. The story is widely shared, and the company in question promises to address the issue. Then either the system is superseded by something new, or technical interventions are made in the attempt to produce results with greater parity. Those results and technical fixes remain proprietary and secret, and the public is told to rest assured that the malady of bias has been "cured."[20] It is much rarer to have a public debate about *why* these forms of bias and discrimination frequently recur and whether more fundamental problems are at work than simply an inadequate underlying dataset or a poorly designed algorithm.

One of the more vivid examples of bias in action comes from an insider account at Amazon. In 2014, the company decided to experiment with automating the process of recommending and hiring workers. If automation had worked to drive profits in product recommendation and warehouse organization, it could, the logic went, make hiring more efficient. In the words of one engineer, "They literally wanted it to be an engine where I'm going to give you 100 resumes, it will spit out the top five, and we'll hire those."[21] The machine learning system was designed to rank people on a scale of one to five, mirroring Amazon's system of product ratings. To build the underlying model, Amazon's engineers used a dataset of ten years' worth of résumés from fellow employees and then trained a statistical model on fifty thousand terms that appeared in those résumés. Quickly, the system began to assign less importance to commonly used engineering terms, like programming languages, because everyone listed them in their job histories. Instead, the models began valuing more subtle cues that recurred on successful applications. A strong prefer-

ence emerged for particular verbs. The examples the engineers mentioned were "executed" and "captured."[22]

Recruiters starting using the system as a supplement to their usual practices.[23] Soon enough, a serious problem emerged: the system wasn't recommending women. It was actively downgrading résumés from candidates who attended women's colleges, along with any résumés that even included the word "women." Even after editing the system to remove the influence of explicit references to gender, the biases remained. Proxies for hegemonic masculinity continued to emerge in the gendered use of language itself. The model was biased against women not just as a category but against commonly gendered forms of speech.

Inadvertently, Amazon had created a diagnostic tool. The vast majority of engineers hired by Amazon over ten years had been men, so the models they created, which were trained on the successful résumés of men, had learned to recommend men for future hiring. The employment practices of the past and present were shaping the hiring tools for the future. Amazon's system unexpectedly revealed the ways bias already existed, from the way masculinity is encoded in language, in résumés, and in the company itself. The tool was an intensification of the existing dynamics of Amazon and highlighted the lack of diversity across the AI industry past and present.[24]

Amazon ultimately shut down its hiring experiment. But the scale of the bias problem goes much deeper than a single system or failed approach. The AI industry has traditionally understood the problem of bias as though it is a bug to be fixed rather than a feature of classification itself. The result has been a focus on adjusting technical systems to produce greater quantitative parity across disparate groups, which, as we'll see, has created its own problems.

Understanding the relation between bias and classifica-

tion requires going beyond an analysis of the production of knowledge—such as determining whether a dataset is biased or unbiased—and, instead, looking at the mechanics of knowledge construction itself, what sociologist Karin Knorr Cetina calls the "epistemic machinery."[25] To see that requires observing how patterns of inequality across history shape access to resources and opportunities, which in turn shape data. That data is then extracted for use in technical systems for classification and pattern recognition, which produces results that are perceived to be somehow objective. The result is a statistical ouroboros: a self-reinforcing discrimination machine that amplifies social inequalities under the guise of technical neutrality.

The Limits of Debiasing Systems

To better understand the limitations of analyzing AI bias, we can look to the attempts to fix it. In 2019, IBM tried to respond to concerns about bias in its AI systems by creating what the company described as a more "inclusive" dataset called Diversity in Faces (DiF).[26] DiF was part of an industry response to the groundbreaking work released a year earlier by researchers Joy Buolamwini and Timnit Gebru that had demonstrated that several facial recognition systems—including those by IBM, Microsoft, and Amazon—had far greater error rates for people with darker skin, particularly women.[27] As a result, efforts were ongoing inside all three companies to show progress on rectifying the problem.

"We expect face recognition to work accurately for each of us," the IBM researchers wrote, but the only way that the "challenge of diversity could be solved" would be to build "a data set comprised from the face of every person in the world."[28] IBM's researchers decided to draw on a preexisting dataset of a hundred million images taken from Flickr, the largest publicly

available collection on the internet at the time.[29] They then used one million photos as a small sample and measured the craniofacial distances between landmarks in each face: eyes, nasal width, lip height, brow height, and so on. Like Morton measuring skulls, the IBM researchers sought to assign cranial measures and create categories of difference.

The IBM team claimed that their goal was to increase diversity of facial recognition data. Though well intentioned, the classifications they used reveal the politics of what diversity meant in this context. For example, to label the gender and age of a face, the team tasked crowdworkers to make subjective annotations, using the restrictive model of binary gender. Anyone who seemed to fall outside of this binary was removed from the dataset. IBM's vision of diversity emphasized the expansive options for cranial orbit height and nose bridges but discounted the existence of trans or gender nonbinary people. "Fairness" was reduced to meaning higher accuracy rates for machine-led facial recognition, and "diversity" referred to a wider range of faces to train the model. Craniometric analysis functions like a bait and switch, ultimately depoliticizing the idea of diversity and replacing it with a focus on *variation*. Designers get to decide what the variables are and how people are allocated to categories. Again, the practice of classification is centralizing power: the power to decide which differences make a difference.

IBM's researchers go on to state an even more problematic conclusion: "Aspects of our heritage—including race, ethnicity, culture, geography—and our individual identity—age, gender and visible forms of self-expression—are reflected in our faces."[30] This claim goes against decades of research that has challenged the idea that race, gender, and identity are biological categories at all but are better understood as politically, culturally, and socially constructed.[31] Embedding identity

claims in technical systems as though they are facts observable from the face is an example of what Simone Browne calls "digital epidermalization," the imposition of race on the body. Browne defines this as the exercise of power when the disembodied gaze of surveillance technologies "do the work of alienating the subject by producing a 'truth' about the body and one's identity (or identities) despite the subject's claims."[32]

The foundational problems with IBM's approach to classifying diversity grow out of this kind of centralized production of identity, led by the machine learning techniques that were available to the team. Skin color detection is done because it can be, not because it says anything about race or produces a deeper cultural understanding. Similarly, the use of cranial measurement is done because it is a method that *can* be done with machine learning. The affordances of the tools become the horizon of truth. The capacity to deploy cranial measurements and digital epidermalization at scale drives a desire to find meaning in these approaches, even if this method has nothing to do with culture, heritage, or diversity. They are used to increase a problematic understanding of accuracy. Technical claims about accuracy and performance are commonly shot through with political choices about categories and norms but are rarely acknowledged as such.[33] These approaches are grounded in an ideological premise of biology as destiny, where our faces become our fate.

The Many Definitions of Bias

Since antiquity, the act of classification has been aligned with power. In theology, the ability to name and divide things was a divine act of God. The word "category" comes from the Ancient Greek *katēgoríā,* formed from two roots: *kata* (against) and *agoreuo* (speaking in public). In Greek, the word can be

either a logical assertion or an accusation in a trial—alluding to both scientific and legal methods of categorization.

The historical lineage of "bias" as a term is much more recent. It first appears in fourteenth-century geometry, where it refers to an oblique or diagonal line. By the sixteenth century, it had acquired something like its current popular meaning, of "undue prejudice." By the 1900s, "bias" had developed a more technical meaning in statistics, where it refers to systematic differences between a sample and population, when the sample is not truly reflective of the whole.[34] It is from this statistical tradition that the machine learning field draws its understanding of bias, where it relates to a set of other concepts: generalization, classification, and variance.

Machine learning systems are designed to be able to generalize from a large training set of examples and to correctly classify new observations not included in the training datasets.[35] In other words, machine learning systems can perform a type of induction, learning from specific examples (such as past résumés of job applicants) in order to decide which data points to look for in new examples (such as word groupings in résumés from new applicants). In such cases, the term "bias" refers to a type of error that can occur during this predictive process of generalization—namely, a systematic or consistently reproduced classification error that the system exhibits when presented with new examples. This type of bias is often contrasted with another type of generalization error, variance, which refers to an algorithm's sensitivity to differences in training data. A model with high bias and low variance may be underfitting the data—failing to capture all of its significant features or signals. Alternatively, a model with high variance and low bias may be overfitting the data—building a model too close to the training data so that it potentially captures "noise" in addition to the data's significant features.[36]

Outside of machine learning, "bias" has many other meanings. For instance, in law, bias refers to a preconceived notion or opinion, a judgment based on prejudices, as opposed to a decision come to from the impartial evaluation of the facts of a case.[37] In psychology, Amos Tversky and Daniel Kahneman study "cognitive biases," or the ways in which human judgments deviate systematically from probabilistic expectations.[38] More recent research on implicit biases emphasizes the ways that unconscious attitudes and stereotypes "produce behaviors that diverge from a person's avowed or endorsed beliefs or principles."[39] Here bias is not simply a type of technical error; it also opens onto human beliefs, stereotypes, or forms of discrimination. These definitional distinctions limit the utility of "bias" as a term, especially when used by practitioners from different disciplines.

Technical designs can certainly be improved to better account for how their systems produce skews and discriminatory results. But the harder questions of why AI systems perpetuate forms of inequity are commonly skipped over in the rush to arrive at narrow technical solutions of statistical bias as though that is a sufficient remedy for deeper structural problems. There has been a general failure to address the ways in which the instruments of knowledge in AI reflect and serve the incentives of a wider extractive economy. What remains is a persistent asymmetry of power, where technical systems maintain and extend structural inequality, regardless of the intention of the designers.

Every dataset used to train machine learning systems, whether in the context of supervised or unsupervised machine learning, whether seen to be technically biased or not, contains a worldview. To create a training set is to take an almost infinitely complex and varied world and fix it into taxonomies composed of discrete classifications of individual data points,

a process that requires inherently political, cultural, and social choices. By paying attention to these classifications, we can glimpse the various forms of power that are built into the architectures of AI world-building.

Training Sets as Classification Engines: The Case of ImageNet

In the last chapter we looked at the history of ImageNet and how this benchmark training set has influenced computer vision research since its creation in 2009. By taking a closer look at ImageNet's structure, we can begin to see how the dataset is ordered and its underlying logic for mapping the world of objects. ImageNet's structure is labyrinthine, vast, and filled with curiosities. The underlying semantic structure of ImageNet was imported from WordNet, a database of word classifications first developed at Princeton University's Cognitive Science Laboratory in 1985 and funded by the U.S. Office of Naval Research.[40] WordNet was conceived as a machine-readable dictionary, where users would search on the basis of semantic rather than alphabetic similarity. It became a vital source for the fields of computational linguistics and natural language processing. The WordNet team collected as many words as they could, starting with the Brown Corpus, a collection of one million words compiled in the 1960s.[41] The words in the Brown Corpus came from newspapers and a ramshackle collection of books including *New Methods of Parapsychology, The Family Fallout Shelter,* and *Who Rules the Marriage Bed?*[42]

WordNet attempts to organize the entire English language into synonym sets, or synsets. The ImageNet researchers selected only nouns, with the idea that nouns are things that pictures can represent—and that would be sufficient to train machines to automatically recognize objects. So Image-

Net's taxonomy is organized according to a nested hierarchy derived from WordNet, in which each synset represents a distinct concept, with synonyms grouped together (for example, "auto" and "car" are treated as belonging to the same set). The hierarchy moves from more general concepts to more specific ones. For example, the concept "chair" is found under artifact → furnishing → furniture → seat → chair. This classification system unsurprisingly evokes many prior taxonomical ranks, from the Linnaean system of biological classification to the ordering of books in libraries.

But the first indication of the true strangeness of ImageNet's worldview is its nine top-level categories that it drew from WordNet: plant, geological formation, natural object, sport, artifact, fungus, person, animal, and miscellaneous. These are curious categories into which all else must be ordered. Below that, it spawns into thousands of strange and specific nested classes, into which millions of images are housed like Russian dolls. There are categories for apples, apple butter, apple dumplings, apple geraniums, apple jelly, apple juice, apple maggots, apple rust, apple trees, apple turnovers, apple carts, 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. It is a riot of words, ordered into strange categories like those from Jorge Luis Borges's mythical encyclopedia.[43] At the level of images, it looks like madness. Some images are high-resolution stock photography, others are blurry phone photographs in poor lighting. Some are photos of children. Others are stills from pornography. Some are cartoons. There are pin-ups, religious icons, famous politicians, Hollywood celebrities, and Italian comedians. It veers wildly from the professional to the amateur, the sacred to the profane.

Human classifications are a good place to see these poli-

tics of classification at work. In ImageNet 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." There is an implicit assumption here that only "male" and "female" bodies are recognized as "natural." There is an ImageNet category for the term "Hermaphrodite," but it is situated within the branch Person → Sensualist → Bisexual alongside the categories "Pseudohermaphrodite" and "Switch Hitter."[44]

Even before we look at the more controversial categories within ImageNet, we can see the politics of this classificatory scheme. The decisions to classify gender in this way are also naturalizing gender as a biological construct, which is binary, and transgender or gender nonbinary people are either nonexistent or placed under categories of sexuality.[45] Of course, this is not a novel approach. The classification hierarchy of gender and sexuality in ImageNet recalls earlier harmful forms of categorization, such as the classification of homosexuality as a mental disorder in the *Diagnostic and Statistical Manual*.[46] This deeply damaging categorization was used to justify subjecting people to repressive so-called therapies, and it took years of activism before the American Psychiatric Association removed it in 1973.[47]

Reducing humans into binary gender categories and rendering transgender people invisible or "deviant" are common features of classification schemes in machine learning. Os Keyes's study of automatic gender detection in facial recognition shows that almost 95 percent of papers in the field treat gender as binary, with the majority describing gender as immutable and physiological.[48] While some might respond that this can be easily remedied by creating more categories, this

fails to address the deeper harm of allocating people into gender or race categories without their input or consent. This practice has a long history. Administrative systems for centuries have sought to make humans legible by applying fixed labels and definite properties. The work of essentializing and ordering on the basis of biology or culture has long been used to justify forms of violence and oppression.

While these classifying logics are treated as though they are natural and fixed, they are moving targets: not only do they affect the people being classified, but how they impact people in turn changes the classifications themselves. Hacking calls this the "looping effect," produced when the sciences engage in "making up people."[49] Bowker and Star also underscore that once classifications of people are constructed, they can stabilize a contested political category in ways that are difficult to see.[50] They become taken for granted unless they are actively resisted. We see this phenomenon in the AI field when highly influential infrastructures and training datasets pass as purely technical, whereas in fact they contain political interventions within their taxonomies: they naturalize a particular ordering of the world which produces effects that are seen to justify their original ordering.

The Power to Define "Person"

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 the 21,841 categories that were originally in the ImageNet hierarchy, noun classes 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 a more *nouny* noun

than the concept of "light," which in turn is more nouny than a concept such as "health."[51] Nouns occupy various places on an axis from the concrete to the abstract, 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. While this approach has the aesthetics of objectivity, it is nonetheless a profoundly ideological exercise.

For a decade, ImageNet contained 2,832 subcategories under the top-level category "Person." The subcategory with the most associated pictures was "gal" (with 1,664 images) followed by "grandfather" (1,662), "dad" (1,643), and chief executive officer (1,614—most of them male). With these highly populated categories, we can already begin to see the outlines of a worldview. ImageNet contains a profusion of classificatory categories, including ones for race, age, nationality, profession, economic status, behavior, character, and even morality.

There are many problems with the way ImageNet's taxonomy purports to classify photos of people with the logics of object recognition. Even though its creators removed some explicitly offensive synsets in 2009, categories remained for racial and national identities including Alaska Native, Anglo-American, Black, Black African, Black Woman (but not White Woman), Latin American, Mexican American, Nicaraguan, Pakistani, South American Indian, Spanish American, Texan, Uzbek, White, and Zulu. To present these as logical categories of organizing people is already troubling, even before they are used to classify people based on their appearance. Other people are labeled by careers or hobbies: there are Boy Scouts, cheerleaders, cognitive neuroscientists, hairdressers, intelligence analysts, mythologists, retailers, retirees, and so on. The existence of these categories suggests that people can be visually ordered according to their profession, in a way that seems reminiscent

of such children's books as Richard Scarry's *What Do People Do All Day?* ImageNet also contains categories that make no sense whatsoever for image classification such as Debtor, Boss, Acquaintance, Brother, and Color-Blind Person. These are all nonvisual concepts that describe a relationship, be it to other people, to a financial system, or to the visual field itself. The dataset reifies these categories and connects them to images, so that similar images can be "recognized" by future systems.

Many truly offensive and harmful categories hid in the depths of ImageNet's Person categories. Some classifications were misogynist, racist, ageist, and ableist. The list includes Bad Person, Call Girl, Closet Queen, Codger, Convict, Crazy, Deadeye, Drug Addict, Failure, Flop, Fucker, Hypocrite, Jezebel, Kleptomaniac, Loser, Melancholic, Nonperson, Pervert, Prima Donna, Schizophrenic, Second-Rater, Slut, Spastic, Spinster, Streetwalker, Stud, Tosser, Unskilled Person, Wanton, Waverer, and Wimp. Insults, racist slurs, and moral judgments abound.

These offensive terms remained in ImageNet for ten years. Because ImageNet was typically used for object recognition—with "object" broadly defined—the specific Person category was rarely discussed at technical conferences, nor did it receive much public attention until the ImageNet Roulette project went viral in 2019: led by the artist Trevor Paglen, the project included an app that allowed people to upload images to see how they would be classified based on ImageNet's Person categories.[52] This focused considerable media attention on the influential collection's longtime inclusion of racist and sexist terms. The creators of ImageNet published a paper shortly afterward titled "Toward Fairer Datasets" that sought to "remove unsafe synsets." They asked twelve graduate students to flag any categories that seemed unsafe because they were either "inherently offensive" (for example, containing profanity or "racial or gender slurs") or "sensitive" (not inherently offen-

sive but terms that "may cause offense when applied inappropriately, such as the classification of people based on sexual orientation and religion").[53] While this project sought to assess the offensiveness of ImageNet's categories by asking graduate students, the authors nonetheless continue to support the automated classification of people based on photographs despite the notable problems.

The ImageNet team ultimately removed 1,593 of 2,832 of the People categories—roughly 56 percent—deeming them "unsafe," along with the associated 600,040 images. The remaining half-million images were "temporarily deemed safe."[54] But what constitutes *safe* when it comes to classifying people? The focus on the hateful categories is not wrong, but it avoids addressing questions about the workings of the larger system. The entire taxonomy of ImageNet reveals the complexities and dangers of human classification. While terms like "microeconomist" or "basketball player" may initially seem less concerning than the use of labels like "spastic," "unskilled person," "mulatto," or "redneck," when we look at the people who are labeled in these categories we see many assumptions and stereotypes, including race, gender, age, and ability. In the metaphysics of ImageNet, there are separate image categories for "assistant professor" and "associate professor"—as though once someone gets a promotion, her or his biometric profile would reflect the change in rank.

In fact, there are no neutral categories in ImageNet, because the selection of images always interacts with the meaning of words. The politics are baked into the classificatory logic, even when the words aren't offensive. ImageNet is a lesson, in this sense, of what happens when people are categorized like objects. But this practice has only become more common in recent years, often inside the tech companies. The classification schemes used in companies like Facebook are much harder

to investigate and criticize: proprietary systems offer few ways for outsiders to probe or audit how images are ordered or interpreted.

Then there is the issue of where the images in ImageNet's Person categories come from. As we saw in the last chapter, ImageNet's creators harvested images en masse from image search engines like Google, extracted people's selfies and vacation photos without their knowledge, and then paid Mechanical Turk workers to label and repackage them. All the skews and biases in how search engines return results are then informing the subsequent technical systems that scrape and label them. Low-paid crowdworkers are given the impossible task of making sense of the images at the rate of fifty per minute and fitting them into categories based on WordNet sysnets and Wikipedia definitions.[55] Perhaps it is no surprise that when we investigate the bedrock layer of these labeled images, we find that they are beset with stereotypes, errors, and absurdities. A woman lying on a beach towel is a "kleptomaniac," a teenager in a sports jersey is labeled a "loser," and an image of the actor Sigourney Weaver appears, classified as a "hermaphrodite."

Images—like all forms of data—are laden with all sorts of potential meanings, irresolvable questions, and contradictions. In trying to resolve these ambiguities, ImageNet's labels compress and simplify complexity. The focus on making training sets "fairer" by deleting offensive terms fails to contend with the power dynamics of classification and precludes a more thorough assessment of the underlying logics. Even if the worst examples are fixed, the approach is still fundamentally built on an extractive relationship with data that is divorced from the people and places from whence it came. Then it is rendered through a technical worldview that seeks to fuse together a form of singular objectivity from what are complex and varied cultural materials. The worldview of ImageNet is

not unusual in this sense. In fact, it is typical of many AI training datasets, and it reveals many of the problems of top-down schemes that flatten complex social, cultural, political, and historical relations into quantifiable entities. This phenomenon is perhaps most obvious and insidious when it comes to the widespread efforts to classify people by race and gender in technical systems.

Constructing Race and Gender

By focusing on classification in AI, we can trace the ways that gender, race, and sexuality are falsely assumed to be natural, fixed, and detectable biological categories. Surveillance scholar Simone Browne observes, "There is a certain assumption with these technologies that categories of gender identity and race are clear cut, that a machine can be programmed to assign gender categories or determine what bodies and body parts should signify."[56] Indeed, the idea that race and gender can be automatically detectable in machine learning is treated as an assumed fact and rarely questioned by the technical disciplines, despite the profound political problems this presents.[57]

The UTKFace dataset (produced by a group at the University of Tennessee at Knoxville), for example, consists of more than twenty thousand images of faces with annotations for age, gender, and race.[58] The dataset's authors state that the dataset can be used for a variety of tasks, including automated face detection, age estimation, and age progression. The annotations for each image include an estimated age for each person, expressed in years from zero to 116. Gender is a forced binary: either zero for male or one for female. Second, race is categorized into five classes: White, Black, Asian, Indian, and Others. The politics of gender and race here are as obvious as they are harmful. Yet these kinds of dangerously reductive cate-

gorizations are widely used across many human-classifying training sets and have been part of the AI production pipelines for years.

UTKFace's narrow classificatory schema echoes the problematic racial classifications of the twentieth century, such as South Africa's apartheid system. As Bowker and Star have detailed, the South African government passed legislation in the 1950s that created a crude racial classification scheme to divide citizens into the categories of "Europeans, Asiatics, persons of mixed race or coloureds, and 'natives' or pure-blooded individuals of the Bantu race."[59] This racist legal regime governed people's lives, overwhelmingly those of Black South Africans whose movements were restricted and who were forcibly removed from their land. The politics of racial classification extended into the most intimate parts of people's lives. Interracial sexuality was forbidden, leading to more than 11,500 convictions by 1980, mostly of nonwhite women.[60] The complex centralized database for these classifications was designed and maintained by IBM, but the firm often had to rearrange the system and reclassify people, because in practice there were no singular pure racial categories.[61]

Above all, these systems of classification have caused enormous harm to people, and the concept of a pure "race" signifier has always been in dispute. In her writing about race, Donna Haraway observes, "In these taxonomies, which are, after all, little machines for clarifying and separating categories, the entity that always eluded the classifier was simple: race itself. The pure Type, which animated dreams, sciences, and terrors, kept slipping through, and endlessly multiplying, all the typological taxonomies."[62] Yet in dataset taxonomies, and in the machine learning systems that train on them, the myth of the pure type has emerged once more, claiming the authority of science. In an article on the dangers of facial

recognition, media scholar Luke Stark notes that "by introducing a variety of classifying logics that either reify existing racial categories or produce new ones, the automated pattern-generating logics of facial recognition systems both reproduce systemic inequality and exacerbate it."[63]

Some machine learning methods go beyond predicting age, gender, and race. There have been highly publicized efforts to detect sexuality from photographs on dating sites and criminality based on headshots from drivers' licenses.[64] These approaches are deeply problematic for many reasons, not least of which is that characteristics such as "criminality"—like race and gender—are profoundly relational, socially determined categories. These are not inherent features that are fixed; they are contextual and shifting depending on time and place. To make such predictions, machine learning systems are seeking to classify entirely relational things into fixed categories and are rightly critiqued as scientifically and ethically problematic.[65]

Machine learning systems are, in a very real way, *constructing* race and gender: they are defining the world within the terms they have set, and this has long-lasting ramifications for the people who are classified. When such systems are hailed as scientific innovations for predicting identities and future actions, this erases the technical frailties of how the systems were built, the priorities of why they were designed, and the many political processes of categorization that shape them. Disability scholars have long pointed to the ways in which so-called normal bodies are classified and how that has worked to stigmatize difference.[66] As one report notes, the history of disability itself is a "story of the ways in which various systems of classification (i.e., medical, scientific, legal) interface with social institutions and their articulations of power and knowledge."[67] At multiple levels, the act of defining categories and

ideas of normalcy creates an outside: forms of abnormality, difference, and otherness. Technical systems are making political and normative interventions when they give names to something as dynamic and relational as personal identity, and they commonly do so using a reductive set of possibilities of what it is to be human. That restricts the range of how people are understood and can represent themselves, and it narrows the horizon of recognizable identities.

As Ian Hacking observes, classifying people is an imperial imperative: subjects were classified by empires when they were conquered, and then they were ordered into "a kind of people" by institutions and experts.[68] These acts of naming were assertions of power and colonial control, and the negative effects of those classifications can outlast the empires themselves. Classifications are technologies that produce and limit ways of knowing, and they are built into the logics of AI.

The Limits of Measurement

So what is to be done? If so much of the classificatory strata in training data and technical systems are forms of power and politics represented as objective measurement, how should we go about redressing this? How should system designers account for, in some cases, slavery, oppression, and hundreds of years of discrimination against some groups to the benefit of others? In other words, how should AI systems make representations of the social?

Making these choices about which information feeds AI systems to produce new classifications is a powerful moment of decision making: but who gets to choose and on what basis? The problem for computer science is that justice in AI systems will never be something that can be coded or computed. It requires a shift to assessing systems beyond opti-

mization metrics and statistical parity and an understanding of where the frameworks of mathematics and engineering are causing the problems. This also means understanding how AI systems interact with data, workers, the environment, and the individuals whose lives will be affected by its use and deciding where AI should not be used.

Bowker and Star conclude that the sheer density of the collisions of classification schemes calls for a new kind of approach, a sensitivity to the "topography of things such as the distribution of ambiguity; the fluid dynamics of how classification systems meet up—a plate tectonics rather than static geology."[69] But it also requires attending to the uneven allocations of advantage and suffering, for "how these choices are made, and how we may think about that invisible matching process, is at the core of the ethical project."[70] Nonconsensual classifications present serious risks, as do normative assumptions about identity, yet these practices have become standard. That must change.

In this chapter we've seen how classificatory infrastructures contain gaps and contradictions: they necessarily reduce complexity, and they remove significant context, in order to make the world more computable. But they also proliferate in machine learning platforms in what Umberto Eco called "chaotic enumeration."[71] At a certain level of granularity, like and unlike things become sufficiently commensurate so that their similarities and differences are machine readable, yet in actuality their characteristics are uncontainable. Here, the issues go far beyond whether something is classified wrong or classified right. We are seeing strange, unpredictable twists as machine categories and people interact and change each other, as they try to find legibility in the shifting terrain, to fit the right categories and be spiked into the most lucrative feeds. In a machine learning landscape, these questions are no less urgent

because they are hard to see. What is at stake is not just a historical curiosity or the odd feeling of a mismatch between the dotted-outline profiles we may glimpse in our platforms and feeds. Each and every classification has its consequence.

The histories of classification show us that the most harmful forms of human categorization—from the Apartheid system to the pathologization of homosexuality—did not simply fade away under the light of scientific research and ethical critique. Rather, change also required political organizing, sustained protest, and public campaigning over many years. Classificatory schemas enact and support the structures of power that formed them, and these do not shift without considerable effort. In Frederick Douglass's words, "Power concedes nothing without a demand. It never did and it never will."[72] Within the invisible regimes of classification in machine learning, it is harder to make demands and oppose their internal logics.

The training sets that are made public—such as ImageNet, UTKFace, and DiF—give us some insight into the kinds of categorizations that are propagating across industrial AI systems and research practices. But the truly massive engines of classification are the ones being operated at a global scale by private technology companies, including Facebook, Google, TikTok, and Baidu. These companies operate with little oversight into how they categorize and target users, and they fail to offer meaningful avenues for public contestation. When the matching processes of AI are truly hidden and people are kept unaware of why or how they receive forms of advantage or disadvantage, a collective political response is needed—even as it becomes more difficult.