2

STRAIGHT CODE

LENNA AND THE ORIGINAL SIN OF COMPUTER VISION

In 1973, Alexander Sawchuk, the father of the JPEG image file format, was an electrical engineer working in the University of Southern California's Signal and Image Processing Institute (SIPI). Sawchuk was looking for the perfect image to scan to optimize the new image compression algorithms that SIPI was developing. He wanted an image that was glossy, had a complex mix of colors and textures, and one that contained a human face. The engineers at SIPI came across a *Playboy* centerfold of Swedish model Lena Söderberg—her name in the magazine spelled 'Lenna' to encourage its proper pronunciation. Lenna wore a feathered Panama hat, boots, stockings, and a pink boa, which seemed to offer the required image properties for testing their compression algorithms. The SIPI engineers took the top third of the centerfold only so that the image would be sized appropriately to be wrapped around the drum of their Muirhead wirephoto scanner and so that the resulting digital image would be 512px by 512px square. The scanner had custom analog-to-digital converters installed to capture the red, green, and blue channels of the scan in digital code that was then stored on a Hewlett Packard 2100 minicomputer.[1]

Emily Chang has referred to this moment as "tech's original sin."[2] By 1991, SIPI had made their scanned image of Lenna available for free to researchers across the world. It had quickly become the standard for evaluating image compression algorithms and could be frequently seen in the pages of image processing journals, books, and conference papers.[3] As Mar Hicks, historian of technology and author of *Programmed Inequality*, told *WIRED* magazine, "If they hadn't used a *Playboy* centerfold, they almost certainly would have used another picture of a pretty white woman. The *Playboy* thing gets our attention, but really what it's about is this world-building that's gone on in computing from the beginning—it's about building worlds for certain people

Figure 2.1
Lenna's Playboy centerfold scan by SIPI. Retrieved from https://en.wikipedia.org/wiki/File:Lenna_(test_image).png.

and not for others."[4] Hicks's comment captures well a sentiment that seems widespread among women working in computer science and engineering.

In 1997, Sunny Bains, a prominent scientist, tech journalist, and editor of engineering journals, wrote an op-ed for *Electronic Engineering Times* in which she argued that "the Lenna image grates because of its exclusivity. It's not difficult to feel isolated when you're a woman working in a male-dominated field. Seeing provocative images of women in learned journals can add to that feeling of non-inclusion."[5] This feeling has endured over time. In 2015, Maddie Zug published an op-ed in the *Washington Post* arguing that the use of the image in computer science curriculum led to sexual comments from male classmates and indicated a broader cultural problem that is at least partly responsible for the depressed numbers of women working in advanced computer science labs.[6] In 2013, Deanna Needell and Rachel Ward published a paper in which they used an image of the Italian-American model Fabio in place of Lenna for image compression research in hopes of motivating their field to reconsider the use of Lenna.[7] Jeff Seideman, an industry leader in image encoding, captured these critiques perfectly in his *defense* of the continued use of the Lenna image, telling the *Atlantic* in 2016 that "when you use a picture like that for so long, it's not a person anymore; it's just pixels."[8]

The use of the Lenna image fits into a long series of literal objectifications of women that have been central to the development of technology, ranging

from the metaphorical objectification of the original labor of women computer operators whose function was automated by increasingly sophisticated circuitry to the literal objectification of women ranging from Kodak's use of "Shirley cards" to optimize their film and film processing technologies to the unauthorized use of Suzanne Vega's voice to perfect the sound compression algorithms that led to the MP3.[9] For nearly fifty years, Lenna has served as the benchmark of image-processing quality, shaping everything from the development of image compression formats like JPEG to the operations of smartphone cameras like Apple's iPhone to the operations of image software like Google Images.

The omnipresent use of the Lenna image is indicative of the unvoiced heteronormativity that permeates Silicon Valley. It harkens back to an earlier détente in the war on pornography in which *Playboy* was allowed to publish objectifications of a particular variety of female bodies and increasingly granted public legitimacy, while such open representations of alternative forms of desire—for different shapes, sizes, anatomies, and colors of bodies, perhaps in different contexts, performing different erotic acts, and so on—were denied such legitimacy and public visibility. It is the assumption of banality, the presumption that such an image was by default uncontroversial, that belies its heteronormativity. As I will show throughout this chapter, this "original sin" can be taken as symbolic of the gender and sexuality-based biases that ground the research and development of new technologies, where similar assumptions of banality, of shared norms, and an expected lack of controversy lead to heteronormative hardware and software.

In particular, we'll look at the history of Google's attempts to automate the censorship of "adult" content via its SafeSearch algorithms and image recognition technologies and Facebook's efforts to streamline the human review of content flagged as inappropriate and produce "human algorithms." While this critique is in no way confined to Google or Facebook—and I intend it to speak to the broader discursive community of computer programmers and software engineers, for which I will use the shorthand "coders"—I will draw heavily on case studies from the two companies to demonstrate the practical effects of this permeation of heteronormativity. The chapter considers this implicit heteronormativity from three perspectives: (1) its permeation into the discursive community of coders themselves, (2) its subsequent permeation into the parameters of the algorithms and datasets that currently shape computer vision as a field, and (3) its ongoing maintenance

by "human algorithms," the people charged with performing the human labor of reviewing content flagged by the system for violating community standards. Across these three domains, we can see that heteronormative biases have a strong impact on the research, development, implementation, and everyday operation of content moderation algorithms.

THE HETERONORMATIVITY OF CODERS

In her article "Going to Work in Mommy's Basement," Sarah Sharma draws on the common Silicon Valley trope of "beta" coders whose conditions of existence are founded upon taking advantage of the unrecognized and feminized labor of their mommies, a twenty-first-century twist on the devaluation and rendering invisible of feminized reproductive and affective labor. She asks, "What kind of work is done in this 'coder's cave' of antisocial techbro culture? What kind of world gets programmed from a position of uncomplicated safety and abundance?"[10] This best of all possible worlds for male coders is what Emily Chang calls a *brotopia*.[11] In this brotopia, men who often identify as spurned lovers or borderline incels in their youth are finally recognized, courted by large tech companies, put in charge of cutting-edge start-ups, and through their power, prestige, and wealth can finally make up for lost time when it comes to sex. Sarah Banet-Weiser has described this as "toxic geek masculinity" and shown that it is not an isolated phenomenon but is instead undergirded by and connected to the broader cultural context of misogyny and heteronormativity online (examined in chapter 1).[12]

Toxic geeks understand themselves as being the victims of marginalization and alpha-male masculinity. Nathan Ensmenger has shown that the tech bros and toxic geeks referred to here are usually shaped by the historical injury of having been geeks, nerds, and socially awkward in their formative years.[13] As Kristina Bell, Christopher Kampe, and Nicholas Taylor explain, they thus understand themselves through the stereotype of being "weak, easily bullied, and socially awkward males who lack social skills, athletic abilities, and physical attractiveness," with their sole redeeming feature and claim to political, economic, and sexual agency being their "perceived [. . .] mastery over digital technologies."[14] Adrienne Shaw argues that because of this felt sense of victimhood, toxic geeks react hostilely to anyone who calls them out as being the perpetrators of abuses of power themselves. They seem totally incapable of recognizing their own privilege and in response receive feminist

critiques as unwarranted attacks, even going so far as to define their identity as anti-feminist.[15] They are thus doubly injured by women, first through sexual rejection and second by feminist critique and women seeking entry into the workplace at technology companies. As Banet-Weiser notes, "This assemblage of features—technological prowess, social awkwardness, and cognitive dissonance about privilege—yields a contradictory subjectivity. According to this frame, geek men have been injured by the world and, more importantly, by women. The aggressive and violent regulation and exclusion of women is a way to regain masculine capacity."[16]

Sue Decker, former president of Yahoo, has used the metaphor of a fish being the last to discover water to describe the ubiquity of gender bias and heteronormative sexual harassment in Silicon Valley.[17] This bears out in what little comprehensive survey data we have from tech companies. Take, for instance, the infamous "Elephant in the Valley" study from 2017, which surveyed women of various ages and ranks that worked in tech companies about their experiences with sexism in the workplace. The study found that 90 percent of women surveyed had experienced sexist behavior at company off-sites or at industry conferences. Further, 60 percent of them had received unwanted sexual advances; most reported these advances were not one-time instances but instead repeated overtures, and more than half came from a superior at their company. A majority of those who reported sexual harassment were dissatisfied with how the company handled their case, and many ended up signing nondisparagement agreements to keep them from going public with their stories. Nearly 40 percent of women who experienced sexual harassment declined to report it for fear it would stunt their career advancement.[18]

This harassment takes place in both the materialized utopias of tech campuses and after work at off-site company events and industry conferences. Tech campuses are built to accommodate frat-like behaviors and to offer all the comforts of "mommy's basement." Most of them offer unlimited free alcohol and games like table tennis and foosball. They regularly keep free high-end food within fifty yards of every employee at all times and offer free dinners for employees who stay after 5 p.m. They contain services on-site ranging from gyms to doctors to hairdressers to laundry to pet care. All of this takes place within open floor plans that make it notably difficult for employees to avoid coworkers who might harass them. In short, their designs skew toward the desires of young, single men. This is perhaps nowhere more

visible than in Apple's failure to include a daycare service in its new $5 billion Apple Park campus that opened in 2017.[19] As Emily Chang has found, "Few employers offer stipends for child care, and even fewer provide on-site child care. Sure, you can bring your dog to work, but you are (mostly) on your own with your baby."[20]

Silicon Valley tries to position itself as being on the cutting edge of both technological and sexual experimentation, with strong polyamorous communities and hookup culture buoyed by exclusive company sex parties hosted at private homes. As Chang has found, most of these events skew toward the fantasies of heterosexual men, as they are maintained with higher ratios of women (selected for their appearance) to encourage sexual encounters with tech bros and toxic geeks. While the Valley's progressivism extends to threesomes, these are almost exclusively a man and two women, with gay and bisexual sex acts conspicuously absent from the scene and little pressure on men to engage in this sort of progressive experimentation. In explaining his peers' behavior, Evan Williams, a cofounder of Twitter, has described polyamory as a "hack."[21] Thus, most of the rhetoric surrounding sex in the Valley is simply a convenient means for justifying the voracious and heteronormative sexual appetites of men who are finally able to get access to women's bodies in the ways they dreamed of as deprived adolescents.[22]

The liberation of this "progressive" scene is exclusively male. Women who participate in sexual exploration lose credibility and respect. They also gain a reputation of being open to any and all future advances, anywhere, and at any time. However, not attending has similarly bad consequences, as it can severely limit women's opportunities to network and advance their careers since work gets done at these sex parties.[23] The women at these parties are also kept at arm's length for fear that they might be "founder hounders," the Silicon Valley neologism for gold diggers. The rhetoric surrounding founder hounders is frequently used to justify predatory behavior toward these women, as it presumes that they are similarly engaging in predatory behavior by trying to trap rich men and extract capital from them. In a chilling interview, Chang spoke to an anonymous tech company founder about the rampant use of drugs to "lubricate" sex parties and the potential advantage tech bros were taking of women. He replied that "on the contrary, it's women who are taking advantage of him and his tribe, preying on them for their money."[24]

A culture like this was able to emerge because women's participation in the field significantly diminished leading up to the dot-com boom and tech's resurgence after the dot-com collapse. This was a particularly notable turnaround when it came to the development of software, which was dominated by women for many decades.[25] While in the early 1980s women were earning nearly 40 percent of all computer science degrees in the United States, that number decreased to closer to 20 percent by the time today's platforms were emerging and has remained relatively stable since. At companies like Google and Facebook, from what numbers are publicly available, women account for between 30 and 35 percent of the workforce, but only around 20 percent of the technical jobs.[26] This lack of representation is particularly acute in AI fields, where 80 percent of professors are men, as are 85 to 90 percent of the research staff at Google and Facebook.[27] During their formative years, many such companies employed aptitude tests like the IBM Programmer Aptitude Test and the Cannon-Perry Test that were biased toward the selection of antisocial, combative, and hubristic coders who just so happened to also be predominantly male. These tests included "brain teasers" that asked applicants to make wild speculations on the spot backed by some form of logic and calculation, like asking applicants how many windows are in New York City. Google, for instance, did not stop using these sorts of brainteasers until 2013. Its longtime former head of human resources, Laszlo Bock, then admitted to the *New York Times* that "brainteasers are a complete waste of time. . . . They don't predict anything."[28]

While companies began to wake up to this problem in the 2010s, much of their culture, corporate policies, and technological infrastructures had already been determined by largely male coding and legal teams. Companies like Google espoused a commitment to hiring more women early on, but this commitment was often half-hearted, as the company's organizational chart reads more like a soap opera script of interoffice affairs. CEO Eric Schmidt, cofounder Sergey Brin, and Andy Rubin, the lead technician who developed Android, all engaged in relationships with women at the company who were their subordinates, and longtime executive Amit Singhal was given a golden parachute after sexually harassing a woman.[29] Despite this bad corporate behavior, the company did strive to implement fairer hiring practices. In 2008, Google established a secret hiring practice in which female applicants had their applications submitted to a second review committee called

the "Revisit Committee" if the initial hiring committee found them unacceptable. The Revisit Committee was tasked with reviewing the applications of all potential diversity hires. Company policy stipulated that hiring committees remain silent about any interviews they conducted. Google also established a secret policy that all technical candidates' committees contain at least one woman, a practice that put undue burden on women already at the company.[30] This intense secrecy and the measures Google took to correct for bad hiring practices demonstrate a key antagonism within Silicon Valley that persists to this day: the antagonism between the myth of meritocracy and the use of hiring practices meant to combat unconscious bias.

Meritocracy may be the central myth around which Silicon Valley's culture is constructed. The problem with this is that belief in meritocracy most often requires a belief that brilliance is innate, and research shows that these cultural biases lead gatekeepers like teachers and hiring committees to assume that (white) men are more likely to possess innate talent. One university study found that "the extent to which practitioners of a discipline believe that success depends on sheer brilliance is a strong predictor of women's and African American's representation in that discipline."[31] Another empirical study found that "when an organization is explicitly presented as meritocratic, individuals in managerial positions favor a male employee over an equally qualified female employee."[32] The problem with meritocracy is that it doesn't recognize the cultural contexts within which "brilliance" is defined and emerges. In Silicon Valley, brilliance is defined in such a way that it privileges male coders, and the position of privilege from which male coders apply to jobs goes unrecognized in the application process. Even Michael Young, who brought the term into public discourse with his 1958 book *The Rise of Meritocracy*, recognized this problem.[33] He concluded that meritocracy could produce a new social stratification and sense of moral exceptionalism based on who had access to elite education and social networks. Further, meritocracy is always impossible to implement because it first needs to be defined, and the definition of meritocracy is most frequently founded on preferences for certain qualities, aptitudes, demeanors, and skill sets that are primarily available to wealthy white men.

True believers in meritocracy don't see these internal contradictions and instead use meritocracy as a logical explanation for the privilege that they enjoy. It gives them a smugness and overinflated sense of self-worth that can cause them to react violently to what they perceive as "discriminatory

affirmative action" policies like the ones Google implemented to hire more women for technical positions. While incurring these violent backlashes may be worth it if diversity hiring actually leads to more equity in the workforce, this doesn't seem to be the case as the number of women in technical positions at technology companies has remained rather stagnant despite the past decade of attempts at fairer hiring practices. Most companies now implement some equivalent of unconscious bias training where they offer employees workshops on how their unconscious biases about race and gender might impact their thinking in the workplace, a new and revised version of earlier attempts at "sensitivity training." There is another problem, however, with how unconscious bias training actually plays out. In attempts to avoid shutting down dialogue by calling employees out on biased behavior, unconscious bias training begins with the premise that everyone has biases, that there is nothing wrong with having biases, and all one is responsible for is curbing them as much as possible. Studies have found that this essentially normalizes gender and racial bias by removing the cultural stigma around it. It can even cause people to accept these biases as unavoidable and make them more likely to exhibit these types of biases in the workplace.[34] Even Anthony Greenwald, the inventor of the Implicit Association Test that helps demonstrate to people the unconscious biases they hold, has expressed concern about unconscious bias training. He told an interviewer, "Understanding implicit bias does not actually provide you with the tools to do something about it."[35]

In short, much of the workforce that is charged with creating the algorithms that govern the internet hold heteronormative biases about gender and sexuality. They often come from a position of privilege, desiring to work from mommy's basement without recognizing the care and benefits that position gives them in the supposed meritocracy they believe themselves to be navigating. Their ideology tends toward the biologization of talent and the belief that brilliance is innate to individual coders. No other explanation could justify the hubris necessary to believe themselves as the ordained arbiters of the future. Further, much of this connects to an understanding of themselves as being ignored by the world, and women in particular, in their adolescence, as they were forced into the position of "betas" or beta males. They have pulled themselves up by their bootstraps and are ready for their just rewards after having proven the world's evaluation of them wrong. Those who disagree with this position are often structurally located in weaker positions in the corporate organizational chart and have

little power to challenge the dominant culture of the valley. All of this looks eerily similar to the worldview espoused by the alt-right, particularly their anxieties around gender and sexuality. Nowhere is this clearer than in the case of the Google memo, to which we'll now turn.

JAMES DAMORE'S GOOGLE MEMO

The most infamous instance of the penetration of heteronormativity, misogyny, and contemporary alt-right ideology into Silicon Valley is easily James Damore's Google memo.[36] In 2017, Damore circulated a memo titled "Google's Ideological Echo Chamber: How Bias Clouds Our Thinking about Diversity and Inclusion" internally within the company that was quickly leaked to the press and became a media sensation. The memo is couched within the framework of human biodiversity, a hobby horse often used by alt-right writers to leverage the authority of scientific objectivity to support their arguments but which tends to produce politically motivated pseudo-scientific arguments. Damore begins the memo by writing, "I value diversity and inclusion, am not denying that sexism exists, and don't endorse using stereotypes. When addressing the gap in representation in the population, we need to look at population level differences in distributions."[37] According to Damore, men and women—N.B., he exclusively uses these terms cisnormatively—differ biologically at the statistical level of population. These differences include

- women being more open to feelings and aesthetics than ideas,
- women having a stronger interest in people than objects,
- women expressing extroversion through gregariousness rather than assertiveness, and
- women being more susceptible to "neuroticism," including having higher anxiety and lower stress tolerance.

For Damore, these differences explain the distribution of men and women into different professions, the gender pay gap, and the retention problem that tech companies have with female employees.

Damore is careful to note that while these biological differences hold at the population level, they do not map directly onto individual men and women. He further outlines some potentially useful "non-discriminatory ways to reduce the gender gap," such as making software engineering more

people-oriented through pair programming and collaboration initiatives, making tech and leadership roles less stressful, and better facilitating work-life balance through options like part-time work. However, Damore's memo is more famous for its other suggestions that echo familiar cries of "reverse racism." Damore argues that it is discriminatory to foster diversity through

- diversity initiatives that offer programs, mentoring, and classes exclusively for women;
- using high priority queues and secondary reviews for female applicants;
- applying advanced scrutiny to groups of people not sufficiently diverse; and
- setting organizational-level objectives and key results for increased representation.

He follows these arguments with suggestions that Google de-moralize diversity, stop alienating conservatives, de-emphasize empathy ("being emotionally unengaged helps us better reason about the facts"), punish intentional sexism rather than unintentional transgressions and microaggressions (he argues here that there is no evidence that speech constitutes violence), be more open about the science of human biodiversity (e.g., IQ and anatomical sex differences), and reconsider making unconscious bias training mandatory for promotion committees, among other things.[38]

Damore describes Google as an "ideological echo chamber" with "extreme" and "authoritarian" elements. He argues that Google—and here he is referring specifically to the midlevel managerial and public relations teams instituting diversity initiatives—is "extreme" in its belief that representational disparities are due to structural injustice. Google is "authoritarian" because it engages in "discrimination"—or what critics like Damore often refer to as "reverse discrimination"—when it tries to institute policies to correct for structural injustice.[39] He understands Google as a "silent, psychologically unsafe environment" that has been invaded by the culture of "PC-authoritarians" (i.e., politically correct authoritarians).[40] Damore noted that he had received "many" messages from supporters within the company who thanked him for raising these issues and who noted that they would have been too afraid to speak out within the company.[41] Thus, Damore understood himself to be standing up for the voiceless inside the company and as taking an acknowledged risk in circulating the memo. Screenshots of

Google's internal message boards, interviews with employees, and an informal Twitter poll all showed that a significant number of Google employees agreed with the contents of Damore's memo.[42] At one point, the document was inaccessible because so many employees were attempting to view it concurrently.[43]

Many women who have worked or currently work for Google have spoken out since the memo to argue that Damore's ideas are endemic to the company. Kelly Ellis, a former Google employee who reported being sexually harassed at the company in 2015, noted that this rhetoric was common at Google, not just among coders but also among those doing performance reviews and on hiring committees.[44] She told *WIRED* that "Those guys like to pretend they're silenced and afraid, but they're not."[45] Another Google employee noted that the response to the memo inside Google was highly gendered, with men being much more likely to agree with Damore and see him as brave for speaking out.[46] A third Google employee noted of the memo, "It's not worth thinking about this as an isolated incident and instead a manifestation of what ails all of Silicon Valley."[47] Megan Smith, a former vice president at Google who also served as chief technology officer for the United States under Barack Obama, similarly noted that these perspectives are common across Silicon Valley and permeate its culture.[48]

If one were inclined to take this evidence as anecdotal, one could look to the 2017 lawsuit in which the US Department of Labor sued Google for the release of decades of employment data in an effort to combat gender bias within the company.[49] Janette Wipper, a Department of Labor regional director, testified in court that "we found systemic compensation disparities against women pretty much across the entire workforce."[50] Janet Herold, the regional solicitor for the Department of Labor, further noted, "The government's analysis at this point indicates that discrimination against women in Google is quite extreme, even in this industry."[51] The lawsuit against Google, in addition to a handful of other Department of Labor suits against Silicon Valley tech companies, was grounded on the fact that these companies were federal contractors.

Two months after they were filed, President Trump signed an executive order that effectively rolled back Obama-era protections for female workers.[52] It is worth noting that as of 2019, the Department of Labor has lost its lawsuit suing for the requisite data to demonstrate a long-term trend of gender bias within Google.[53] While this story was largely passed over silently

in the press, Trump's executive order is reflective of repeated libertarian arguments that the gender pay gap is a myth and alt-right arguments that any gender pay gap is due to human biodiversity rather than cultural bias and structural injustice. Google made similar claims that it had closed the gender pay gap at all levels across the entire company when it refused to hand over the additional data that the Department of Labor requested. Thus, the context within which Damore wrote was one in which a number of female Google employees at both junior and senior levels were accusing the company of frequently harboring similar sexist beliefs and in which the best data available to the Department of Labor led them to believe there was a systemic gender pay gap across the entire company.

Damore was fired shortly after the memo was leaked—although a Harvard-Harris poll would show that 55 percent of surveyed voters said that Google was wrong to fire Damore.[54] In his op-ed for the *Wall Street Journal*, Damore described Google as seeking to placate the outraged mob that resulted from his memo being leaked. He wrote, "The mob would have set upon anyone who openly agreed with me or even tolerated my views."[55] Key Google executives and other Silicon Valley elites voiced their condemnation of the memo, including Danielle Brown, Google's VP of diversity; Sundar Pichai, Google's CEO; Sheryl Sandberg, COO of Facebook; Susan Wojcicki, CEO of YouTube; and Megan Smith, a former Google VP. However, even Pichai, the ultimate authority at Google, equivocated in his statement, writing, "[T]o suggest a group of our colleagues have traits that make them less biologically suited to that work is offensive and not OK. [. . .] At the same time, there are co-workers who are questioning whether they can safely express their views in the workplace (especially those with a minority viewpoint). They too feel under threat, and that is also not OK."[56] This equivocation remains to this day, as Google has barred its employees from protesting the company's actions in their official capacity as employees or anywhere near Google's Pride Parade float at the 2019 San Francisco Pride Parade.[57] In response, a number of employees have petitioned the San Francisco Pride board of directors to revoke Google's sponsorship of the 2019 Pride Parade.[58]

Following his firing, Damore mounted a publicity campaign in which he began to increasingly echo the public's interpretation of his message, quickly dropping his caveats about applying population statistics to individuals and his potentially helpful suggestions for reform and instead focusing on ramping up his image as a victim and his insistence on human biodiversity as a

central cause of gender disparities in the workplace. In his Reddit Ask Me Anything (AMA), Damore noted, "I honestly haven't seen any valid criticism that disputes my claims."[59] Damore's positioning of himself as a Silicon Valley pariah has led to his adoption by the alt-right in North America.[60] This is perhaps nowhere more evident than in Damore's photoshoot with Peter Duke, who the *New York Times* has described as "the Annie Leibovitz of the alt-right."[61] In the resulting photo, Damore sits in a T-shirt that reads "Gulag" styled as the Google logo. Damore arranged to have this photo retweeted by Mike Cernovich, an alt-right conspiracy theorist who has previously claimed that date rape does not exist. Afterward, Damore claimed that he was unaware of Cernovich's politics and past statements and only did it to reach Cernovich's 300,000 followers.[62] Like many like-minded coders in Silicon Valley, Damore keeps his politics hard to pin down, hiding behind claims of ignorance, claims of centrism, reliance on the rhetoric of science, and caveats about his potentially having some form of undiagnosed autism as an excuse for any insensitivity in his statements. On his AMA, Damore described himself as "centrist" and a "liberal," but in a group for libertarian-leaning Google employees, he more accurately noted that his libertarianism "influenced a lot of the document."[63]

In his AMA, Damore also noted that a key influence on his thinking was University of Toronto pop psychologist Jordan Peterson, who blends vague and thus easily universalizable morals with antiquated Jungian analytical psychology and highly motivated readings of empirical evidence of the biological differences of anatomical sex. In the wake of the media campaign, Peterson interviewed Damore for his YouTube channel, ostensibly to provide an objective assessment of the Google memo. Despite Peterson's claims to scientific objectivity, he found nearly all Damore's ideas to be well supported by "the relevant psychological science." Peterson argues that Damore, in fact, holds what is the majority viewpoint and that Damore was only silenced and made to feel like a pariah because "social constructionists" are better organized—despite their being wrong factually, scientifically, and ethically. Peterson even describes affirmative action hiring practices as "racist." The result is a revivified Damore, who in the end argues that he has been proven right, that the entire culture is attempting to silence any dissenting viewpoints, and that we need a more "objective" way of looking at these issues.[64]

It is worth noting that others who have fact-checked Damore's memo have had very different takes and have found the scientific evidence for many of his

claims to be either totally lacking or in contradiction to his statements.[65] Anatomical sex differences actually don't hold much explanatory power when it comes to people's different abilities, attitudes, and actions.[66] In a survey of nearly four thousand studies, boys do not perform better than girls at mathematics as children, and the advantages adolescent and adult men have in mathematical ability are much better explained by social conditioning and cultural biases.[67] And while differences in anatomical sex do correlate to different occupational interests—like an interest in STEM careers—these differences are not biological. They are much more likely because of the discourse in communities surrounding different occupations, as well as social conditioning.[68] These differences are exacerbated when it comes to working with computers, as it has long been known that males exhibit "greater sex-role stereotyping of computers, higher computer self-efficacy, and more positive affect about computers."[69] This is not only the consensus among researchers doing empirical studies of the very issues that Damore raises but also the standard position of the American Psychological Association.[70] As Diane Halpern, professor of psychology and past president of the American Psychological Association, has noted, the problem comes when these differences are understood as deficiencies and interpreted as biologically preordained, when in fact they result from a complex and continuous feedback loop between biology and environment.[71]

What we can learn from this is that while Google increasingly seeks to diversify its labor pool and offer a voice to women at the managerial level, it does not, and likely cannot, fully commit itself to these endeavors. Silencing the discourse on human biodiversity within the company potentially alienates too large a group of the essential talent pool of male coders that the company needs to keep happy in order to operate its global empire. At the top of the pyramid, Sundar Pichai equivocates about his commitment to gender equity in the company, and at the bottom, myriad coders express deep sympathies with Damore's position. This pseudoscientific biologizing of people's abilities, attitudes, and actions according to anatomical sex is not only inaccurate and reductive of the complexities of anatomical sex but also erases the hard-earned and central distinction between anatomical sex and gender. This erasure leads to a slippage in which gender roles are easily essentialized through the same pseudoscientific appeals to biology. By combining sex and gender, gender also becomes binarized. This cisnormativity, as we've seen, undergirds heteronormativity. It is only atop this cisnormative binarization that

heterosexuality is semicoherent as a concept and available as a cultural norm. Further, it is only atop this binarization that homosexuality can emerge as a derivative and abnormal concept. Instead of individual bodies connected by desire, we have categorically distinct bodies connecting within pre-articulated matrices of desire ([male, female], [male, male], [female, female]).

Coders operating within this epistemological framework are ill-suited to ethically manage the vagaries of contemporary sexuality as it manifests itself through digital communications. And further, because of its pseudoscientific grounding and the increasing retrenchment that occurs after pariahs like Damore are turned into martyrs by alt-right media, we are left with a discourse community surer of its convictions. While we will continue to draw on internal case studies from Google, as we've seen, this conjuncture is in no way limited to a single corporation but instead is endemic to Silicon Valley. This problem is exacerbated by the ambiguous messages of CEOs, the often-toothless warnings of middle managers working toward diversity initiatives, and the very silence that Damore identified in his memo when it comes to internal dialogues about gender equity and diversity. As we will see in the next section of this chapter, when these biases and silences are combined with the hacker culture surrounding the implementation of new algorithms and curation of big data in Silicon Valley, it can lead to biased technological systems and platforms that carry with them a large amount of inertia that inhibits the full correction of biased functions after implementation.

THE HETERONORMATIVITY OF CODE

There is a hubris embedded at the core of Silicon Valley research and development practices that is frequently referred to as the *hacker ethic*. This ethic is unique to the conjuncture in which computer science arose, a cross-fertilization of military and academic research.[72] It was brought to popular awareness by Steven Levy in 1984 when he published *Hackers: Heroes of the Computer Revolution*, which celebrated a culture obsessed with openness, empowerment, and the fundamental maxim that "information wants to be free."[73] Levy's interlocutors made convincing counterarguments at the time, such as Dennis Hayes, who argued that the hacker ethic was a myth constructed by computer journalists and a highly misleading representation of the field. Instead, Hayes saw a culture that was blind to purposes and solely fixated on techniques, a necessity because of its need to bow to corporate and

military priorities to achieve research and development funding. Hackers were so obsessed with manifesting the innovations they envisioned that they were blind to their potential impacts on society.[74] All systems had bugs that could not be predicted. The hacker's job was to build the technology and make ad hoc adjustments to it to fix any errors or ill effects that might emerge. Hackers have a strong confidence that only they can arbitrate the future of technology, and any attempts to regulate them or rein in "progress" are ill-conceived. As Noam Cohen noted, "There is the successful entrepreneur's belief that the disruption that has made him fabulously wealthy must be good for everyone."[75]

Few at the time recognized the gender bias that was being established at the foundation of tech culture. Levy described hackers as so obsessed with programming computers that they would ignore women. He wrote, "Not only an obsession and a lusty pleasure, hacking was a mission. You would hack, and you would live by the Hacker Ethic, and you knew that horribly inefficient and wasteful things like women burned too many cycles, occupied too much memory space."[76] One hacker that Levy quotes uncritically noted, "Women, even today, are considered grossly unpredictable. How can a hacker tolerate such an imperfect being?"[77] Instead, hackers gendered computers and experienced them as their ideal women whose hardware and software could be directly interacted with at will, perfectly controlled, and intimately known. As Noam Cohen has explained, "If this all sounds sort of sexual—or like an old-fashioned marriage—well, you aren't the first to notice."[78] As computer science pioneer John McCarthy noted, "What the user wants is a computer that he can have continuously at his beck and call for long periods of time."[79] The masculine generic in McCarthy's statement is emblematic of a culture that did not forbid women from participating but made a point of not accommodating or welcoming them into the field, increasingly discouraging women from participating in a field they had dominated during its infancy. This effacement of women's historic centrality to computation is deeply connected to the myth of meritocracy in Silicon Valley, as predominantly male, libertarian individualists continually perpetuate a narrative in which they arrive at fame and fortune without having had any special privileges or owing anything to anybody.[80]

This hacker ethic quickly cemented itself into what others have called "the Californian Ideology," an aggressive libertarian and narcissistic understanding of society that masquerades under the façade of chill nerds who just like

to build cool things.[81] Whether this belief system is maintained in earnest by all programmers in the Valley is irrelevant. As scholars like Christian Fuchs and Nick Dyer-Witheford have pointed out, programmers are an increasingly precarious class because of their replaceability and are easily controlled by the corporate officers of their companies because of their desire to maintain the perks of their positions—prestige, high wages, utopic office spaces, and the ability to perform labor that they find meaningful.[82] Thus, those programmers who might develop an interest in the purposes of their work or find themselves critical of the social impacts their research might have on the world are left with little room to voice these qualms. Instead, the ruling ideology is one in which "progress"—here understood as advancements in practical technologies—is inevitable, and all one can do is try to capitalize on being the first to meet the bleeding edge of the future. This ideology is established on a fundamental heteronormativity that genders and sexualizes the computer as the perfect object for the masculine gaze and control. The narcissistic hubris that it establishes leads men to believe that no one can see the future better than them, that no one ought to prevent them from realizing their ideas, that any idea will inevitably be made manifest, and that all one is responsible for is hacking together the best operational prototype possible from available resources and patching it as problems emerge in the future. This is precisely the worldview that we will see in Google's development of SafeSearch and Facebook's content moderation practices (Facebook has gone so far as making the address of its campus 1 Hacker Way, Menlo Park, California). Both companies hack together available resources without clear plans or solicitation of outside feedback or criticism. Both companies consider progress to be inevitable and work to be at its cutting edge. And both companies end up embedding heteronormative and sexist bias into the foundations of their platforms that, as we'll see in the following chapters, can never be fully patched after the code has been hacked together.

In the next section, I'd like to turn to a closer examination of the datasets and algorithms behind the automation of content moderation online with a specific focus on Google SafeSearch. While the technical details in the section may be difficult and tedious to some, I think they are worth exploring in this level of detail for a number of reasons. First, if we want to make changes to the algorithms and datasets that shape large portions of the internet, we are going to need to be able to engage in discussions with computer scientists, and this necessitates working toward at least a basic command of their

discourse. It is my hope that going into this level of detail and demonstrating at least a basic awareness of computer science discourse will help make my arguments more convincing to people at the levers of power. Second, I think that these analyses pay dividends, which readers will see if they persist through some of the denser paragraphs. I've done what I can to make things as clear and concise as possible, but the technical literature is dense and difficult to perfectly distill. That said, I've tried to distribute new and surprising findings throughout the extended case study that wouldn't have been possible for me to unearth without diving into this technical literature. Readers can rest assured, though, that the following section of the chapter—and the remainder of the book for that matter—return to less technical issues, like the human labor of content moderation here and the impact that overbroad censorship has on LGBTQIA+ communities in chapters 3 and 4.

GOOGLE SAFESEARCH AND THE CLOUD VISION API

The history of SafeSearch is nearly synonymous with the history of Google. At the turn of the millennium, Google was already more focused on obliging potential advertisers by censoring pornography from its search results than it was on Y2K. One of its earliest hires was Matt Cutts, who for nearly twenty years led the department at Google that fights spam and search engine optimizers to protect the integrity of Google's search results, one of the most important positions at the company. Yet Cutts's first job at the company was to develop SafeSearch. His first months at the company were spent crawling web porn looking for largely text-based classificatory signals that he could use to automate porn filtering and subsequently trying to recruit colleagues to search for porn that might have evaded his filter system.[83] It is worth noting that from the beginning, Google has understood web pornography through the lens of spam. Just like spam, porn has no fixed definition and requires vigilant updates.[84] For Google, porn is like a virus, constantly mutating in form and strategy to evade detection and infect the healthy body of search results.

In its earliest iterations, SafeSearch was focused on Boolean textual analysis almost exclusively. Cutts's web crawlers would analyze the text that appeared on porn sites to aggregate a set of weighted "trigger words" that could indicate the likelihood that any given site was pornography. The viral understanding is evident here, as, for instance, slang and misspellings were considered to be motivated—i.e., deliberate attempts to evade the filter—and

thus were programmed to weigh in as indicators of pornographic content.[85] Google would then layer behavioral data from its users atop this textual data. By keeping track of what users actually clicked on when they were searching for pornography and how long they visited those links, Cutts was able to establish further patterns about what the context and content of websites were.[86] This was particularly important because, at the time, it was impossible to parse the content of images or videos on the web. One could only simulate an understanding of any given image's content through an analysis of the textual content it was embedded in and behavioral data on how users interacted with it. Analyses of images thus began with looking at the text and user behavior attached to them, and only later would these analyses become sophisticated enough to examine the pixel values of the images themselves.

The visual analysis of images pixel by pixel only started to pick up steam in 2008 as graphics processing units (GPUs) became cheaper and more powerful. The first iterations would index the RGB values of millions of images such that any given image could be correlated with nearly identical versions online. This was the origin of the broader capacity we all enjoy today of using an image as a search query on Google, an innovation brought about by Google's focus on porn censorship.[87] Shortly thereafter (c. 2012), Google began exploring the use of machine learning to train neural networks to detect pornographic content and developed what in April of 2016 it would make available to the public as its Cloud Vision API.[88] As Google explains it, "Google Cloud Vision API [application programming interface] enables developers to *understand the content of an image* by encapsulating *powerful machine learning models* in an easy to use REST API" (emphasis mine).[89] Cloud Vision's features include not only "Explicit Content Detection" but also "Label Detection," "Web Detection," "Face Detection," "Logo Detection," "Landmark Detection," "Image Attributes," and "Optical Character Recognition."[90] In my experience working with Cloud Vision, a number of these features remain severely limited, but the API's capacity to detect explicit content is uncannily accurate—provided we understand explicit content as being any and all nudity and that we understand nudity as female-presenting nipples and breasts, genitals, and (sometimes) buttocks.

Google actually has a much larger definition of explicitness that it has programmed into its Cloud Vision API. Images may be considered explicit based on their participation in any of five separate categories: (1) adult, (2) medical, (3) spoof, (4) violent, and (5) "racy" images can all be detected and

blocked. Adult images may contain elements such as nudity, pornographic images or cartoons, or sexual activities.[91] The category is meant to focus solely on "explicit" or "pornographic" nudity, especially those images that focus on "strategic" parts of the anatomy. However, the system is trained to avoid flagging as adult content any medical, scientific, educational, or artistic nudity, as well as "racy" images that cover said "strategic" parts. Medical content consists of "explicit images of surgery, diseases, or body parts," and its classifier primarily searches for "graphic photographs of open wounds, genital close-ups, and egregious disease symptoms." Spoof content primarily looks for memes, which are indicated by the presence of text (often at the top and bottom of images) and typical meme faces, images, and backgrounds. Violent content consists of images flagged as depicting killing, shooting, or blood and gore.[92]

The fifth category was added only after launch and remains in a sort of beta state despite being available to any developers using Google's Cloud Vision API. "Racy" image detection is meant to capture all the content that escapes the adult content filter but might still be risqué enough to be worth censoring. In perhaps the only extant definition of what this content consists of, Google writes, "Racy content includes lewd or provocative poses, sheer or see-through clothing, closeups of sensitive regions, and more."[93] It appears to be most often triggered by images of nudity wherein "strategic" parts are just barely obscured or covered. This is perhaps Google's most nebulous classifier and demonstrates their orientation toward pornography as a virus needing eradication. In this metaphor, the broadness of the classifier indicates that it is more important to eradicate any viral pathogens than it is to preserve benign organisms. In more practical terms, blocking porn is more important than *not blocking* nonporn, including *art*. Take the Venus de Milo, for example. When I ran a Cloud Vision analysis of a standard Wikimedia Commons image of the statue—and keep in mind this is an image Google has certainly indexed, including its surrounding content and context—the API is convinced that it is likely a "racy" image (see figure 2.2).[94]

Before moving on to examine some examples of heteronormative biases that are hardcoded into the datasets that these algorithms are trained on, it is worth outlining some rudimentary results that I obtained by running sets of images through the Cloud Vision API to get a sense of how these sorts of heteronormative biases inflect content moderation on Google's platform. I did a simple Google Image Search for "female breasts" and gathered the first

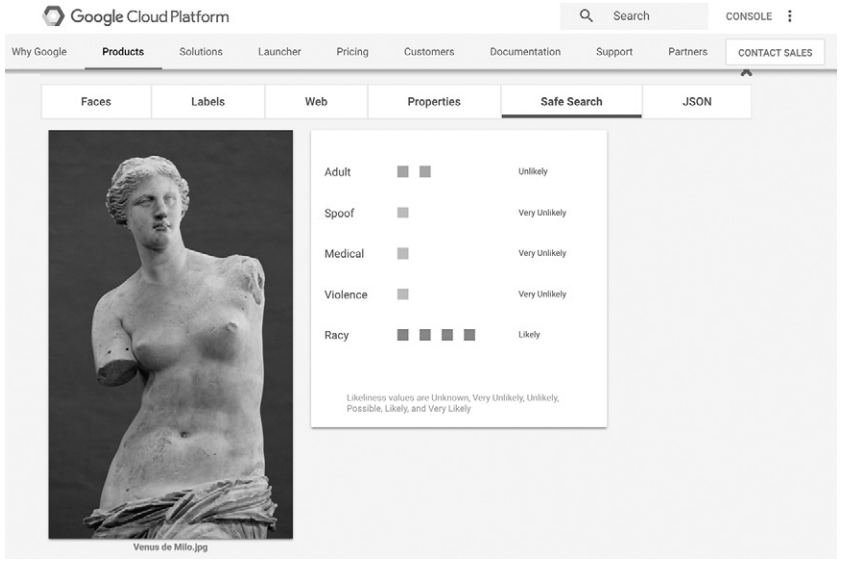


Figure 2.2
Venus de Milo being run through Google's Cloud Vision API.

one hundred relevant images—including a large number of pictures of fully clothed women, medical images and diagrams, and artistic renderings—and ran them through Cloud Vision. Of these images, exactly half of them were determined to "very likely" be "racy" images and thus would be censored in many instances through SafeSearch and in apps developed with the Cloud Vision API. Google SafeSearch seems to have learned the shape and texture of the average female-presenting—and lighter-skinned—breast. This was confirmed by running images of "nude paintings," "nude sculptures," and "hentai" (Japanese-styled nude and sexual drawings) through the system, all of which were frequently flagged as "racy" content when they contained any semblance of a female-presenting breast, again, even when clothed. Needless to say, this result was not repeated when I ran images of bare male-presenting chests through the system.

This betrays a particularly American, heteronormative interpretation of what breast tissue is and what it means. It exacerbates a sexualization of women's and female-presenting bodies that has been a problem for internet users with what platforms deem "female breasts" for decades. For instance, Tarleton Gillespie has excellently documented the decade-long struggle that

people have faced in trying to post images of their breastfeeding online.[95] This problem is hardcoded into the datasets that algorithms like these are trained on, in the first instance by the decision to assume stable gender binaries. These assumptions have been productively challenged by trans women like Courtney Demone, whose #DoIHaveBoobsNow? campaign on Instagram showcased topless photos at different phases of her hormone therapy to beg the question of when her breasts became a content violation.[96] This sexism in the dataset allows for breasts that are coded as "female" to be associated with "pornography," "adult content," or "raciness," thus capturing and reinforcing a culturally singular cisnormative and heteronormative bias. It would take a team of much more capable researchers than me to fully catalogue the results of many of the sexual and gender biases in these datasets. While it is beyond the purview of this book to give a full demonstration of all their impacts, I will now turn to tracing some of the other biased sexual concepts that get captured and reinforced in both the primary datasets that image recognition and computer vision algorithms are trained on and tested against.

At this point, we need to take a detour through how a computer vision algorithm learns to detect adult content so that we can later understand how and where heteronormative biases can be hardcoded into the system. Many machine learning applications require a large dataset with consistent metadata from which they can then analyze and learn patterns to identify and classify new data. In the case of computer vision, this means that large repositories of images must be consistently tagged with appropriate metadata *before* any algorithms can learn to identify and classify new images. Since 2012, ImageNet has been the gold standard image dataset for training computer vision algorithms. ImageNet began as a conference poster presentation by Princeton University researchers in 2009.[97] By 2010, it already contained nearly fifteen million labeled images.[98] In that year, ImageNet also launched the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where computer scientists used a specified subset of the images as seed images to train algorithms to automatically identify and classify images not used in the seed set—you use half the dataset to train your algorithm and then test it on the other half of the images that it hasn't yet analyzed.[99] As we'll see shortly, it was in response to the ILSVRC that the first major breakthrough in the use of convolutional neural networks for computer vision was achieved, and this breakthrough serves as the bedrock for many of Google's computer vision applications today. Further, Google's Inception architecture and

GoogLeNet algorithm were developed atop ImageNet in 2014, later serving as the foundations for Google Photos and the Cloud Vision API.[100]

As noted above, each image in ImageNet needs to be consistently labeled with metadata. The metadata that each of these images can be labeled with, and thus the entire structure of the dataset, is extracted from WordNet, "a large lexical database of English."[101] WordNet also originated at Princeton in 1985 with funding by US Office of Naval Research, the National Science Foundation, the Defense Advanced Research Projects Agency, and the Disruptive Technology Office (formerly the Advanced Research and Development Activity). The goal of WordNet is to capture all of the distinct concepts in the English language and their interrelations. It does this by collecting all English nouns, verbs, adjectives, and adverbs and grouping them into sets of cognitive synonyms that it refers to as "synsets." As its site notes, "Synsets are interlinked by means of conceptual-semantic and lexical relations."[102] Take, for example, the WordNet entry for "sex": WordNet's understanding of sex is composed of four noun synsets, one for "noun.act" that looks at sex as an action and contains "sexual activity" and "sexual practice," one for "noun.group" that looks at anatomical sex, one for "noun.feeling" that looks at sex as an urge, and one for "noun.attribute" that looks at gender and sexuality.[103] The noun.act synset for sex is embedded within the parent synset for a "noun.process" composed of the terms "bodily process," "body process," "body function," and "activity," which themselves are contained within the parent synset "organic process" and "biological process." This latter synset is contained within the "noun.Tops" parent synset of "process" and "physical process" described as "a sustained phenomenon or one marked by gradual changes through a series of states."[104] It is embedded within two more generic noun.Tops synsets, the first being "physical entity," which describes "an entity that has physical existence" and "entity," which describes "that which is perceived or known or inferred to have its own distinct existence (living or nonliving)."[105] In short, WordNet provides the ontology for ImageNet, determining what can exist and how it can be related—with relations existing between parent, child, and sibling concepts.

WordNet's understanding of sex also subsumes the following child synsets:

> "bondage," "outercourse," "safe sex," "conception," "sexual intercourse," "intercourse,' "coitus," "sexual congress," "sexual relation," "relation," "carnal knowledge," "defloration," "fuck," "screw," "ass," "nookie," "piece of tail," "roll in the hay," "shag," "shtup," "hanky panky," [*sic*] "penetration," "unlawful carnal

knowledge," "criminal congress," "extramarital sex," "free love," "adultery," "criminal conversation," "fornication," "incest," "pleasure," "sexual love," "love-making," "love," "carnal abuse," "coupling," "mating," "conjugation," "sexual union," "assortative mating," "disassortative mating," "hybridization," "hybridisation," "crossbreeding," "crossing," "interbreeding," "hybridizing," "dihybrid cross," "monohybrid cross," "reciprocal cross," "reciprocal," "testcross," "test-cross," "inbreeding," "servicing," "service," "reproduction," "procreation," "facts of life," "miscegenation," "crossbreeding," "interbreeding," "generation," "multiplication," "propagation," "biogenesis," "biogeny," "foreplay," "stimulation," "caressing," "cuddling," "hugging," "kissing," "petting," "smooching," "snogging," "feel," "perversion," "sexual perversion," "paraphilia," "exhibitionism," "immodesty," "fetishism," "pedophilia," "paedophilia," "voyeurism," "zoophilia," "zoophilism," "pederasty," "paederasty," "sodomy," "buggery," "anal sex," "anal intercourse," "oral sex," "cunnilingus," "cunnilinctus," "fellatio," "fellation," "cock sucking," "blowjob," "soixante-neuf," "sixty-nine," "autoeroticism," "autoerotism," "masturbation," "onanism," "self-stimulation," "self-abuse," "frottage," "jacking off," "jerking off," "hand job," "wank," "promiscuity," "promiscuousness," "sleeping around," "one-night stand," "lechery," "homosexuality," "homosexualism," "homoeroticism," "queerness," "inversion," "sexual inversion," "lesbianism," "sapphism," "tribadism," "bisexuality," "straightness," "bestiality," and "zooerastia."[106]

While it is hard to keep a data structure like this in your head—and I'd recommend taking a look at the term "sex" and others via WordNet's online platform to get a better sense of it—it is clear that WordNet is engaging in some pretty sophisticated ontological work. It essentially offers an entire linguistic and conceptual schematization of the world ready-made and in machine-readable form.

As Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan have shown, "language itself contains recoverable and accurate imprints of our historic biases, whether these are morally neutral as towards insects or flowers, problematic as towards race or gender, or even simply veridical, reflecting the *status quo* for the distribution of gender with respect to careers or first names. These regularities are captured by machine learning along with the rest of semantics."[107] This is certainly the case with WordNet, where we can find several standard conceptual biases about anatomical sex, gender, and sexuality embedded in the English-language semantics that are formalized in the synsets connected to "sex." For example, one synset for masturbation combines "self-stimulation" with "self-abuse," both defining "manual

stimulation of your own genital organ for sexual pleasure."[108] Here we can see the theological concept of "onanism" go digital (see the introduction). Historical heteronormative biases surrounding masturbation and procreative sex are rendered in machine-readable form.

The term "sodomy" is found in two child synsets for the term "sex." The first synset also contains buggery, anal sex, and anal intercourse, while the second also contains bestiality and zooerastia.[109] Thus, sodomy forms a machine-readable bridge between anal sex with humans and sex with animals, a common trope in conservative fearmongering that surfaces in many debates surrounding LGBTQIA+ rights. It was in fact precisely these connections that Justice Antonin Scalia drew upon in his dissent in *Lawrence v. Texas* (2003), the case that legalized gay and lesbian sex in the United States.[110] The terms "crossbreeding" and "interbreeding" are found in two child synests for the term "sex." The first synset also contains the terms "hybridization," "hybridisation," "crossing," and "hybridizing" and is defined as "(genetics) the act of mixing different species or varieties of animals or plants and thus to produce hybrids."[111] The second synset also contains the term "miscegenation" and is defined as "reproduction by parents of different races (especially by white and non-white persons)."[112] Thus we can see not only the biological essentialism of sexuality that is a hallmark of heteronormativity but also the continued life of scientific racism in machine-readable form. Thomas F. Gossett has catalogued the United States' long legacy of besmirching scientific discourse by leveraging its ethos to peddle scientifically incorrect conflations of race and species. However, while Gossett hoped that Franz Boas, among others, largely delegitimated such nonsense in at least scientific if not popular discourse, here we can see it manifesting once again in a foundational dataset for computer science in the twenty-first century.[113]

These are just a few of the more glaring biases found in a cursory review of a single search result in the online version of WordNet. Others are certainly waiting to be found. It is unfortunately beyond the purview of this book to extend this analysis much further. However, it is essential that other interested researchers push this work forward by further connecting our legacy of critical and analytical knowledge to the analyses of semantic biases in machine learning platforms that are already being implemented by STEM scholars. Every facet of historical prejudice in English-language discourse is likely to rear its head in machine learning platforms and will largely go unaddressed if none of us are keeping track. Bias and prejudice surrounding sex

and sexuality are perhaps most likely to take center stage, as the centrality of ad revenue—and thus of censoring pornography—will make definitions of sex and sexuality primary foci for machine learning moving forward.

While WordNet contains over 100,000 synsets, ImageNet primarily borrows the nouns, which account for over 80,000 of WordNet's synsets. As they note, "In ImageNet, we aim to provide on average 1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated."[114] Ironically, this Anglocentric dataset is produced by using Amazon's Mechanical Turk to outsource most of the English-language labeling labor to (predominantly) non-native English speakers. Any linguistic barriers are overcome by redundancy and exploitation: have multiple people label the same images, use only the labels that the majority agree on, and only pay those who provided the labels consistent with the majority. The more obfuscated and thus potentially more nefarious problem that may exist here is the influence of Anglocentrism on the deep structure of the datasets. Researchers have pointed out some of the problems in the use of English as the root structure for translation algorithms.[115] To my knowledge, no one has yet sufficiently analyzed what the global effects might be of structuring our computer vision algorithms in accordance with English-language "conceptual-semantic" and "lexical" relations.

Today, ImageNet contains an estimated one hundred million images, and its maintainers hope to expand the dataset to trillions of images in the future.[116] This would make the visual dataset reach a similar scale to the linguistic and conceptual datasets already powering search algorithms—and particularly graph search functions like those at Facebook or Google.[117] This task shines new light on the willingness of companies like Google and Facebook to host infinite and increasingly high-resolution user images for free. And further, labeling the image datasets of the future will likely require some combination of the automation of image labeling, gamifying the practice to stimulate users to perform labeling labor for free, and continuing to hire out the labeling labor through services like Amazon's Mechanical Turk. The economic and political stakes of this increasing emphasis on extracting un- or underpaid labor and producing an objectified, alienating, privately owned, and blackboxed form of collective or social knowledge are already being explored by other scholars.[118]

In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton used ImageNet to make a breakthrough in machine vision at the University of

Toronto when working on an algorithm for the ILSVRC. They used a convolutional neural network (CNN or ConvNet), which is a specific kind of artificial neural network that has the benefits of having fewer connections and parameters and thus being easier to train with only slightly worse performance. The only drawback is that they require large amounts of nonserial or GPU processing power.[119] As Hinton, Sutskever, and Hinton noted, "Their capacity can be controlled by varying their depth and breadth, and they also make strong and mostly correct assumptions about the nature of images (namely, stationarity of statistics and locality of pixel dependencies)."[120] This power is afforded by the ability of CNNs to process convolutional data, which means building a function out of the integration of two other functions or variables. In essence, convolutions are capable of working with the fuzziness of visual data to make accurate identifications. Visual data comes in so many more permutations and positions than the linguistic-based conceptual data that many other artificial neural networks are trained to process. Whereas syntax, grammar, and spelling in English-language textual discourse provide a somewhat standardized conceptual topography (i.e., words are often in the same positions in sentences and rarely misspelled), the visual concept of "cat" as expressed in any given cat image could see that cat positioned in any part of the image, at any distance, in many colors, sizes, positions, fur lengths and textures, with various contexts and backgrounds, and so on.

To provide an unavoidably reductive explanation, Hinton, Sutskever, and Hinton's system was able to do this through the unique feed-forward model of a CNN, which connects three types of layers: (1) convolutional, (2) max pooling, and (3) fully connected (see figure 2.3). In this model, the essential component is the neuron, which is fed pixel values as its inputs and is triggered when it detects a particular pattern in those pixel values (such as a horizontal edge, a vertical edge, or a color contrast).[121] Convolutional layers use many identical copies of these neurons and cluster them together into various kernels that only get triggered when all of the neurons in that kernel are themselves triggered (thus, a kernel might be triggered when it detects a vertical edge, a particular texture, and a particular color contrast all together). Convolutional layers break individual images into small groups of pixels and feed each group of pixels into a set of kernels (itself a set of neurons). In a sense then, in the convolutional layers, the set of kernels the machine has learned range over the image one section at a time and fire when they detect particular patterns (boundaries, textures, shapes, and so on). The data from these convolutional

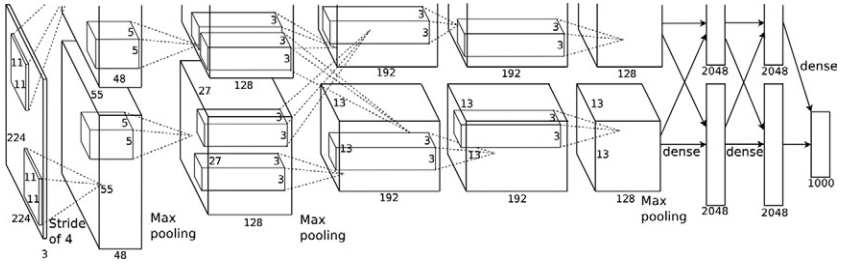


Figure 2.3
Example of layers in a CNN. *Source:* Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems* 25 (2012): 1101.

layers, which essentially consists of which kernels were triggered by which sections of the image being processed, is fed into a max-pooling layer, which can aggregate this local knowledge into a broader regional knowledge about specific types of patterns in the image. A max-pooling layer can thus determine where the edges of objects might be, what textures and colors objects might have, or what shapes they are composed of or contain. This data is then fed into fully connected layers that can create global patterns—and thus global knowledge—of the image based on the regional and local patterns that have been identified. In short, once the max-pooling layer passes on information about patterns across the entire image—like the center of this image has a round figure with a fuzzy texture with two pointy shapes atop it, two circles in it, and lines coming just off-center on either side to its exterior—the fully connected layer can identify this as an image of a cat's face by its shape, fur, pointy ears, round eyes, and whiskers.

The truly unique thing about CNNs though is not their capacity to identify new images but their ability to, when fed a set of seed images with consistent labels, *learn* which kernels of which neurons are useful for indicating which local patterns are useful for indicating which regional patterns, which are useful for identifying global patterns—and thus images themselves. Perhaps more simply, the neurons, kernels, and patterns that the machine uses to identify new images are not *programmed*, they are *learned* by the machine itself through massive and incredibly fast trial-and-error experiments. It is for this reason that CNNs are so hard for people to imagine, as they see images in ways that are very different from us, and that can only be roughly represented to us. For instance, they might identify an image of a cat based on the

texture of their eyeballs, the curvature of their inner ears, the number and placement of whiskers, or even more difficult visual signifiers for humans to distinguish. That said, *what* they can identify is determined in advance by the WordNet labels and the ImageNet images—the algorithm cannot learn to identify things that it does not have images of or labels for. In short, no matter how sophisticated the system, if you feed it heteronormative data, it will produce heteronormative results.

Visual datasets have been shown to contain selection biases that lead to what in computer science lingo are referred to as "certain most discriminative instances" of image categories. We might think of these as certain images that best capture or represent the inherent biases of a particular dataset, and they can demonstrate to both algorithms and even the human eye the differences between visual datasets—things like average depth of focus, number of objects, position of identified objects within the frame, number of identified objects, and so on. For example, in 2011, Antonio Torralba and Alexei A. Efros published the results of some experiments that were inspired by a game called *Name That Dataset!* that they devised in their computer vision lab.[122] In the game, computer scientists working with different visual datasets like ImageNet were presented with three representative images from twelve popular visual datasets and asked if they could match each set of three images to the visual dataset they had been taken from (see figure 2.4). In their lab, most contestants were able to accurately attribute 75 percent of the images to their parent datasets. When they trained an image classifier to play *Name That Dataset!*, the best classifiers were able to achieve 39 percent accuracy—while random chance would have been 8 percent accuracy, thus demonstrating strong evidence of visual biases. Torralba and Efros argued that all datasets are *motivated* because they are pitched, funded, and developed as reactions against the deficits in their predecessors. These motivations tend to lead to biases in the visual data they aggregate, and this bias goes unnoticed because there is very little investigation of cross-dataset generalization. Or, in short, no one is paying much attention to dataset bias.

The biases of the images selected for visual datasets shape everything from the neurons and kernels to the schema and ontology of computer vision platforms. That said, the types of biases implicit in most visual data have less glaring politics and have to do with things like important objects occupying the center and focal middle ground of images, shaping the more banal aspects of what a machine vision system pays attention to. While we certainly need

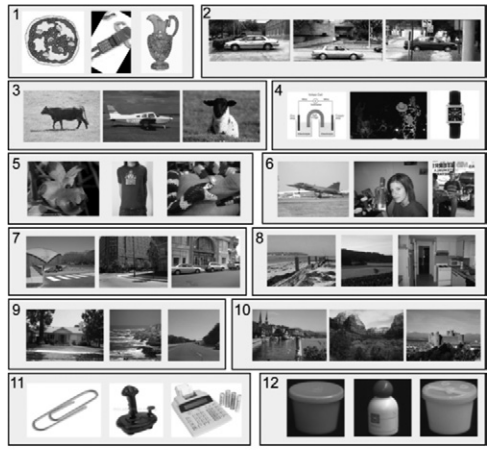


Figure 1. Name That Dataset: Given three images from twelve popular object recognition datasets, can you match the images with the dataset? (answer key below)

Figure 2.4
Name That Dataset! (example of bias in visual datasets). Reproduced from Antonio Torralba and Alexei A. Efros, "An Unbiased Look at Dataset Bias," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, June 20–25 (Piscataway, NJ: IEEE, 2011), 1521.

more humanists contributing to this body of analysis, many of these biases constitute a problem that computer scientists are at least economically motivated to solve because these biases can impact system performance in ways that tech companies care about.

Literally what can exist, how, and where for the system is shaped by these biases, which has repeatedly been demonstrated in terms of race. This was seen most publicly and hauntingly in the 2009 video of an HP facial recognition system failing to register Black Desi and track his face with its webcam.[123] Scholars like Joy Buolamwini and Timnit Gebru, who was unceremoniously fired in December 2020 for raising ethical concerns as co-lead of the Ethical Artificial Intelligence Team at Google, have demonstrated in great detail how racial and gender bias contained in the datasets and parameters of popular facial recognition systems lead to Black women being misclassified up to 34.4 percent of the time.[124] Race is deeply connected with adult content filters, as one of the primary strategies that computer scientists have employed since

at least the 1990s to accurately filter pornographic images has been to focus on detecting skin tones through color and texture profiles.[125] Many scholars have noted the difficulty that these algorithms have in accurately assessing whether nudity is present in images, with error rates routinely ranging from 3 to 10 percent—which means a lot of false positives given that they are analyzing and making decisions about billions of images per day. That said, I've yet to find any scholars performing work like Buolamwini and Gebru's, and most of the assessments of skin tone detection accuracy do not break down error rates by either race/ethnicity or the standard Fitzpatrick skin typology.[126] Only one study examined a specific race/ethnicity in regard to skin tone detection algorithms and found that many systems performed poorly on people from the Indian subcontinent. Their improved system was only able to achieve accuracy rates between 88 and 91 percent.[127]

Based on prior evidence, we can expect that if the datasets these algorithms are trained on do not include representative samples of populations or include biased representations of certain populations the result will be algorithms that exhibit systematic errors when it comes to identifying and classifying POC.[128] It turns out that this is precisely the case with ImageNet. As I've demonstrated elsewhere, the synset on ImageNet that gathers images of Black people consists of images in low resolution that show few facial details, that have bodies positioned further away from the camera, that strongly feature celebrities (around 1 percent of the entire dataset is pictures of Barack Obama) and memes. Most inexcusable, however, is that over 6 percent of the entire category's dataset is composed of images of white people dressed in blackface, largely due to images of Dutch people dressed as Zwarte Piet (i.e., "Black Pete") during their Christmas celebrations.[129] While it is difficult to estimate the impact this might have on adult content filters without more systematic evidence, it is safe to assume that adult content filters will have higher error rates for images of POC and BIPOC women in particular. It is thus likely that POC are experiencing higher false-positive rates where their nonpornographic content is unjustly flagged by automated content moderation systems.

Another major bias that remains is the Western context of these visual datasets, which were largely compiled when images from the US dominated the internet. Google researchers have found that this has led to failures of image recognition systems to accurately identify scenes from other cultural contexts and geographic locations.[130] One of the most frequently

Figure 2.5
Wedding photographs with Google's label predictions based on Open Images. *Source:* Tulsee Doshi, "Introducing the Inclusive Images Competition," *Google AI Blog*, September 6, 2018. https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html.

cited examples is the ability to identify things like weddings, brides, and grooms because of cultural and geographic differences in wedding attire and locations (see figure 2.5). Facebook's AI lab has similarly found that image recognition algorithms also demonstrate embedded cultural biases when they label objects, as they are 15 to 20 percent more likely to incorrectly identify objects from non-Western and low-income communities.[131] It is certainly the case that US- and Western-centric biases about what constitutes obscenity and pornography are embedded in these algorithms as well, as most exposed female-presenting breasts or buttocks and any genitalia will trigger the algorithm globally regardless of whether that particular community would consider that nudity to be pornographic. It will perform even worse at interpreting community standards regarding what constitutes artistic nudity.

Google has increasingly been trying to combat this US and Western bias in its algorithms through its Crowdsource app, which asks users to contribute free labor akin to Amazon's Mechanical Turk with tasks like translation, translation validation, handwriting recognition, sentiment evaluation, and landmark recognition.[132] The company has plans to combat cultural bias in image recognition systems with its Inclusive Images Competition, where it challenges you to use its Open Images dataset to train an algorithm that can successfully be applied to two challenge datasets that Google collected from their global user community via Crowdsource.[133] Of the nearly 35,000 images, fewer than fifty could be described as depicting scantily clad bodies. The most risqué image I found was an outline of Bart Simpson showing his

butt, and more often, the closest thing to racy or risqué images were images of men in tank tops or sleeveless shirts playing basketball. While the Inclusive Images Competition is a worthwhile endeavor, it certainly does not contain the correct images to properly train a machine learning algorithm to make higher-order distinctions about types of nudity based on cultural contexts—like what is artistic and what is culturally normalized versus what is censorable for its prurience.

Another important concern when it comes to the datasets specifically designed to train adult content filters is the consent of the people who are depicted in the images used to train the algorithms. While ImageNet and Open Images are the only publicly accessible image datasets that the image recognition algorithms at Google are known to employ, it is likely that they have propriety datasets in-house for this purpose as well. It is industry practice to ignore concerns over consent when collecting image datasets at this scale, and we might take an example from public adult image datasets used to train algorithms to produce deepfake porn as an example of the issues over consent that arise with these sorts of datasets. After scouring subreddits like r/GeneratedPorn and /AIGeneratedPorn and interviewing coders working on deepfake pornography, Motherboard found that many of these datasets included not only images without people's consent but also images of porn from producers who have been accused of lying to women and coercing them into having sex on camera. These include images from sites like Girls Do Porn, which stands accused of human trafficking and rape. Perhaps most notably, they include images from Czech Casting because each Czech Casting video came with a photoset that was extremely appealing to machine learning programmers. As Samantha Cole explains,

> Each video of a woman also comes with a uniform set of photographs. Each set includes a photograph of the woman holding a yellow sign with a number indicating her episode number, like a mugshot board. Each set also includes photographs of the women posing in a series of dressed and undressed shots on a white background: right side, left side, front, back, as well as extreme close ups of the face, individual nipples, and genitalia.[134]

The obsession with objectification in the mainstream heteroporn industry makes it a particularly appealing sample for adult image datasets, which, coupled with its sheer abundance and availability online, likely ensures that it is strongly over-represented in adult image datasets. Again, without stronger

empirical evidence, it is hard to be certain, but this is a likely explanation for the high incidence of LGBTQIA+ content being unduly filtered by automated content moderation algorithms online that we'll see in chapter 3. Having more mainstream heteroporn in the dataset means not only that it is better at identifying mainstream heteroporn but also that it is better at distinguishing between what is heterosexual porn and what is not heterosexual porn. It is likely less accurate at making the distinction between pornography and nonpornography when it comes to LGBTQIA+ content.

While it is easily imaginable that Google's public relations department would try to externalize the causality of these biases by laying them at the feet of the social collective whose data they mine or the digital laborers working through platforms like Amazon's Mechanical Turk to label the data they train their algorithms on, this clearly is not the case. The meanings established for the dataset's categories prefigure what data will eventually populate them. Take, for example, the term "closet queen," one of three child synsets for the synset of "homosexual," "homophile," "homo," and "gay" in WordNet. A closet queen is defined as "a negative term for a homosexual man who chooses not to reveal his sexual orientation."[135] In its 2011 dataset—the most easily accessible online—ImageNet had thirty-two images representing the term "closet queen" (see figure 2.6). While in its current instantiation, the "closet queen" category is not very threatening and perhaps even laughably bad, it is a very good indicator of the potential implications of such a dataset. Anonymous Mechanical Turk laborers are presented with images of human bodies and prompted to provide this derogatory label to those images based on the presumed sexual identity of the people depicted. The architecture of the dataset demands that stereotypes about what constitutes the successful performance of a particular sex, gender, and sexuality become hardwired into the visual dataset. Regardless of which images end up populating the category, the category's very existence determines the way a computer will see—it will see stereotypically. For example, two men hugging, especially from behind, is a key indicator of closeted homosexuality.

As Alexander Cho has shown, the "default publicness" of social media platforms can lead to LGBTQIA+ youth being outed by computers, which has tragic consequences and reinforces heteronormativity by encouraging youths with unsupportive families or communities to avoid producing or consuming any online content that might out them.[136] This is exacerbated by a system increasingly data mining not only their sexuality but also the sexual semantics

Figure 2.6
Images for "closet queen" synset on ImageNet.

of all the web content they interact with. Beyond this, it is easy to imagine a much more intentional and nefarious future application of such a technology for the automation of outing, where people performing machine-readable acts of closeted queerness become automatically identifiable. While some might view this as the imaginary of dystopic science fiction, I would caution against such a quick dismissal. In 2017, Yilun Wang and Michal Kosinski engineered a deep neural network to analyze images of people's faces and determine their sexual orientation. Their system used publicly available images from a dating site they have refused to name in hopes of slowing copycats.[137] Wang and Kosinski's system was able to accurately distinguish between "gay" and "heterosexual" men in 81 percent of cases and 74 percent of cases in women (compared to human success rates of 61 percent and 54 percent, respectively).[138] While a number of scholars posted critical responses online to the preprint version of the article, demonstrating the limitations of the system,[139] it is hard not to be frightened by the potential capacities of these systems, especially when their visual datasets include contextual data beyond faces (clothes, locations, comportments, other people, and so on), operate at web scale, and incorporate human semantic labeling through Amazon's Mechanical Turk.

The United States and the United Kingdom, in particular, have a long history of selling technology with few to no strings attached to oppressive regimes around the world, ranging from IBM's sale of tabulators to support the Third Reich's "final solution," as documented by Edwin Black, to more recent sales of metadata-based surveillance technologies by Britain's Government Communications Headquarters, an intelligence and security organization, to Honduras, Bahrain, Saudi Arabia, China, and Qatar.[140] And even if Western governments and companies were to exercise a previously unheard of self-restraint by refusing to sell computer vision technologies with such capabilities to regimes interested in the automation of outing, ImageNet is publicly available, as are many of the computer science write-ups of computer vision implementations built atop ImageNet. Anyone, from domestic neo-Nazi alt-right groups to oppressive governments abroad could build such a system themselves were it not available for purchase ready-made, provided ImageNet continues to build out its visual catalogue for terms like "closet queen." Even if this image data is not used to out people, the counting and classifying of LGBTQIA+ people has a long history of rendering them susceptible to dehumanization and violence.[141] This historical LGBTQIA+ precarity is only exacerbated now that private corporations control web-scale data collections and data analytics tools.[142]

In both WordNet and ImageNet, as well as in the image recognition algorithms built atop them, like Google's SafeSearch and Cloud Vision API, we can see the hacker ethic at work. Programmers are focused exclusively on implementing their ideas through the most practical means, largely ignoring the potential social harms these new technologies might cause or assuming that any ill effects can be patched on an ad hoc basis. The datasets that serve as the foundation for the majority of computer vision applications in the world today are riddled with biases, most notably biases about sex, gender, and sexuality. These biases deeply impact how the machine learning algorithms trained on them operate and likely can never be adequately patched after the fact. Biased data will always produce biased results. Without fostering interdisciplinary and diverse dialogue on what unbiased data might look like and large-scale investment in implementing less biased datasets, the infrastructure of the internet will continue to reinforce our preexisting prejudices and further marginalize LGBTQIA+ communities. Lastly, the most common industry response is that human reviewers are the answer for correcting these biases after the fact. However,

as we'll see in the next section and chapter 3, these human reviewers put into practice just as much heteronormative bias as the algorithmic systems they are meant to correct.

THE HETERONORMATIVITY OF CONTENT REVIEW LABOR

FACEBOOK'S "HUMAN ALGORITHMS"

While few humanities and social sciences scholars have unpacked at length the operations of automated content filters, like those discussed above, a number of them have investigated their human counterparts, frequently composed of an underpaid, overburdened, and globalized labor force responsible for censoring broad swaths of the internet.[143] I would contest that this latter phenomenon can best be understood in relation to efforts to automate content moderation through machine learning algorithms like natural language processing systems and computer vision or image recognition systems. The way that major tech companies envision and situate this labor, structure and schematize it, and mask it behind confidentiality agreements and compartmentalization will all strongly reflect these companies' ideas and practices from designing algorithms. In fact, as we'll see, companies like Facebook even describe these laborers as "human algorithms." While the public archive surrounding Google's Cloud Vision API allowed for unique insight into their automated content moderation practices, we will now turn to Facebook's human content moderators because their response to criticism in the wake of the 2016 US election led to them opening up their content moderation practices to the public in unique ways that offer the best insight into how these "human algorithms" are at work within the company.

Facebook only began publishing data on the enforcement of its Community Standards in 2018. In their first report, they found that between seven and nine content views out of every ten thousand were of pieces of content that contained violations of its adult nudity and pornography standards.[144] In 2019, that number was up to eleven to fourteen views per ten thousand.[145] In their latest report, the company notes that since October of 2017, between 0.05 and 0.15 percent of all Facebook content contained flagged violations of the adult nudity and sexual activity clauses of the Community Standards. In each quarter since then, the company has censored between twenty to forty million pieces of content. Around 96 percent of all flagged content was caught by Facebook's automated content moderation system, with the

remaining 4 percent being flagged by the user community.[146] Many of these determinations are considered by the company to be obvious, but the ones that fall into gray areas are kicked up to human reviewers whose labor has been formalized by the company such that they are sometimes referred to as "human algorithms."[147]

The labor force performing these reviews of flagged content is largely hired through a California-based outsourcing firm named oDesk, which farms out content moderation labor for both Google and Facebook, largely hiring from call centers. Around 2012, Facebook employed only fifty moderators for the entire platform, largely from Asia, Africa, and Central America. They were paid $1 per hour plus incentives for reviewing certain amounts of content during their four-hour shifts that could bring their total pay up to $4 an hour—this was the same year Facebook had its initial public offering at $100 billion.[148] In the wake of the 2016 election and Facebook's numerous scandals ranging from Russian trolls to Cambridge Analytica, the company was employing 4,500 content moderators.[149] By 2018, it was employing 7,500 with plans of increasing that number to 15,000.[150] While these numbers have been released, the company maintains secrecy about the number and location of its moderating hubs. As the content moderation labor force has been increased, training has been streamlined. New contract laborers receive two weeks of training and a set of prescriptive manuals for assessing content. They also are given access to Facebook's Single Review Tool (SRT), which allows them to act like human algorithms, categorizing content and checking whether it meets the appropriate sections of Facebook's Community Standards.[151]

These manuals and the SRT are created by young engineers and lawyers at the company who work to distill all content moderation into a series of yes-no decisions, thus producing an algorithm that can be run on the outsourced laborers' bodies and minds. While Facebook claims that there are no time constraints on these laborers, inside information indicates that moderators have eight to ten seconds to review each piece of content (longer for videos), and they have targets of around a thousand pieces of reviewed content per workday. The materials that have been released have all been in English, requiring laborers not fluent in English to use Google Translate throughout their daily work and increasing the difficulty of accurately moderating content.[152] It is worth noting as well that Facebook currently does not have enough training data prepared for its automated content flagging systems to

be very accurate in languages other than English and Portuguese. Despite these linguistic difficulties, moderators are collectively required to review over ten million pieces of content per week and are expected to review every piece of flagged content on the platform within twenty-four hours. The company aims for a benchmark error rate of less than 1 percent, which means that there are still tens of thousands of moderation errors made each day by the platform's human algorithms.[153] As Max Fisher notes, "[M]oderators, at times relying on Google Translate, have mere seconds to recall countless rules and apply them to the hundreds of posts that dash across their screens each day."[154]

A number of these materials have been leaked to the press and can offer a small window into content moderation labor at Facebook. However, as Tarleton Gillespie notes, what is most shocking about the documents is not any aspect in particular that they reveal but the fact that they had to be leaked in the first place. As Gillespie notes,

> These are secret documents, designed not for consideration by users or regulators, but to instruct and manage the 3000+ independently contracted clickworkers who do the actual moderation work. These criteria, while perhaps crafted with input from experts, have not been made public to users, not benefited from public deliberation or even reaction. A single company—in fact a small team within that single company—have anointed themselves the arbiters of what is healthy, fair, harmful, obscene, risky, racist, artistic, intentional, and lewd.[155]

It is precisely at this point of hubris, where a small group of people thought that they could universalize determinations of obscenity in secret, that heteronormativity slipped into the foundation of Facebook's content moderation policies. This bias is only exacerbated in practice as an overworked and underpaid globally distributed set of laborers are charged with implementing them at scale. People performing this labor told the *Guardian* that "moderators often feel overwhelmed by the number of posts they have to review—and they make mistakes, particularly in the complicated area of permissible sexual content."[156]

The problem is that Facebook does not recognize analyzing sexual content as being among the content moderation tasks that are most difficult and time sensitive. The structure of its human algorithm is such that content like hate speech, conspiracy theorists preying on mass shootings, and content the

media is focusing on, like Russian trolls, are "escalated" to better trained and longer-employed laborers. Sexual content more often remains de-escalated, handled by the least trained laborers. Sarah T. Roberts, a professor at UCLA who studies content moderation, told Motherboard, "The fundamental reason for content moderation—its root reason for existing—goes quite simply to the issue of brand protection and liability mitigation for the platform. It is ultimately and fundamentally in the service of the platforms themselves. It's the gatekeeping mechanisms the platforms use to control the nature of the user-generated content that flows over their branded spaces."[157] This is reflected in the training documents and guidelines that have been leaked to the press, which warn moderators against creating "PR fires" by making decisions about content removal that could "have a negative impact on Facebook's reputation or even put the company at legal risk."[158] Heteronormative bias does not often fit the bill for dedicated attention at the company and has never produced serious discussions among politicians about better regulating the platform. The few PR fires that ignite over LGBTQIA+ discrimination are more easily quelled by patronizing apologies and promises of changes and self-regulation that rarely manifest.

While specific examples of biased content moderation decisions will be overviewed in greater detail in chapter 3, it is worth taking a look briefly at two PR fires caused by content moderation decisions at the company to illustrate the point. For example, the first PR fire around heteronormative content moderation on the platform to catch the public's attention occurred in 2011 when Facebook moderators decided to censor an image of a gay kiss taken from the British television drama *Eastenders*. The company apologized profusely and reinstated the image, but there are no available public records or indications within leaked documentation of larger changes being made within its content moderation policies after this encounter.[159] In 2016, Facebook censored a famous image of the so-called napalm girl from the Vietnam War for violating its policies on depicted nudity, thus demonstrating that photorealism is so habitually associated with pornography that all nude bodies are liable to be considered sexually explicit unless proven otherwise after public outcry.[160] This demonstrates the default worldview of Facebook's content moderators in which all female-presenting bodies are sexualized. Facebook similarly apologized after the instance, and it now appears as an example in their training manuals, but this photo is allowed

because of its credentials and historic importance. The next photo like it will still likely be censored on the platform until it has achieved enough awards, hung in enough museums, and climbed the Google Image search rankings.

Thus, Facebook's human algorithm is produced by a small set of predominantly white, straight, young men looking for the most practical solutions imaginable within their normative worldviews that minimally meet the company's desire to protect its brand value and avoid legal liabilities. Its outsourced labor force is made up of culturally heterogeneous contract laborers who receive little training, are given heteronormative guidelines, are under immense pressure to rapidly determine whether to censor content, and are instructed to default to censoring all potentially sexually explicit content, and the most highly trained and accurate of whom dedicate most of their time to reviewing escalated content exclusively. This leads to a lot of heteronormative bias in Facebook's content moderation practices and can have disastrous consequences for its users. Facebook Pages and Groups whose moderators have five pieces of content censored within ninety days or have more than two "elements" that can be considered sexual solicitation or nude imagery on their home pages are unpublished.[161] These policies similarly apply to personal accounts, with many users facing repeated and increasingly lengthy bans from the site. As we'll see in chapter 3, such biased content moderation policies are the norm on platforms like Facebook. They have dire consequences for all users, but particularly for LGBTQIA+ communities, and the adjudication mechanisms for correcting biased decisions are severely lacking. As evidence of this, I'd now like to turn to an examination of Facebook's Community Standards, particularly as they relate to key issues for the LGBTQIA+ community.

FACEBOOK'S COMMUNITY STANDARDS

Facebook maintains a detailed set of community standards and enforcement guidelines for moderating content on their platform. The company describes its mission as including the embrace of diversity in perspectives and notes that because of this, they err on the side of allowing content to persist on the site. This is in large part because Facebook regulates content via a single, global set of rules that are meant to be applied consistently to their entire Facebook community of some two billion users. They further reserve the right to deviate from the letter of these Community Standards and to enforce them

based on the "spirit" of the policies.[162] As we'll see, this leads to a pervasive heteronormativity in Facebook's content moderation practices, despite their best efforts to combat this bias, as well as pervasive Western, and specifically US-centered, biases—a limitation particularly difficult to combat because the deliberations about content moderation policies and the publicly available information on these changes are in English.

Since 2016, Facebook has taken steps to make this deliberative process that results in changes to their Community Standards more transparent to the public. These meetings are prefaced by multiple working groups performing research within the company and engagements in discussions with "stakeholders" who provide input into proposed policy changes, all of which is aggregated and presented at Facebook's Product Policy Forum (previously called the Content Standards Forum) for comment before Facebook makes its final decision on policy changes. Since 2017, Facebook has publicly released the minutes and presentations from the Product Policy Forum on the Facebook Newsroom website, though the stakeholders and forum participants are anonymized in the publicly released documents.[163] A number of people performing content moderation review labor at Facebook have also leaked Facebook's guidelines to the press, particularly from the pre-2017 era in which their work was more thoroughly shrouded in secrecy. From these documents, we can piece together a rather clear picture of how content moderation labor is performed at the company and how the company's Community Standards get formalized into enforcement guidelines that can be run as "human algorithms" on human reviewers.

Currently, Facebook's Community Standards outline twenty-one different types of content that it moderates on its platform. Under the heading "Safety," Facebook's Community Standards list "Sexual Exploitation of Adults" as one type of content that they will moderate on their platform. This includes "content that depicts, threatens or promotes sexual violence, sexual assault, or sexual exploitation," and "content that displays, advocates for, or coordinates sexual acts with nonconsenting parties or commercial sexual services."[164] We can already see here the influence of NCOSE's concept of intersectional sexual exploitation, as sex work is definitionally conflated with sexual exploitation, thus removing agency from sex workers and reinforcing their historic positioning as the corrupted and helpless victims of the darker aspects of society in need of saving. This discourse is rooted in a heteronormative gender binary that historically understands women as

being less compelled to seek sexual pleasure and personally and privately in charge of defending their sexual virtue from nonmonogamous or commercialized sexual activity. While it is certainly noble to guard against real sexual exploitation on a social media platform like Facebook, conflating that task with combating prostitution betrays deeply embedded normative and Christian conservative moral frameworks.

Under the heading of "Objectionable Content," Facebook also lists "Sexual Solicitation." Violations of this portion of the Community Standards include content that "facilitates, encourages or coordinates sexual encounters between adults."[165] This means that users cannot make posts that attempt to coordinate or recruit participants for filmed sexual activities, strip club shows, live sex performances, erotic dances, or sexual, erotic, or tantric massages. Facebook also considers content that explicitly and implicitly solicits sex to violate this standard. Explicit sexual solicitation includes offering or asking for sex or sexual partners, engaging in sex chat or conversations, or sending nude images. Implicit sexual solicitation is defined as an offer or request to engage in sexual activity combined with the use of suggestive statements, sexual hints, shared sexual content, and what Facebook calls "sexualized slang." While these policies officially apply to publicly posted content, Facebook has a history of scanning private messages as well and unclear policies about when and how it does so.[166] Essentially, Facebook does not want anyone to use its platform to arrange sex acts of any kind, including digital sex acts like sexting. As journalist Violet Blue notes, "Anything encouraging sex for pleasure between adults is now a bannable offense in public posts."[167] This anti-sex approach to content moderation should not be surprising, given Facebook's normative moral stance on pornography and sex writ large. In fact, Facebook was one of the first major tech companies to break ranks and support the passage of FOSTA-SESTA by the US Congress, a sweeping anti-sex and anti-pornography bill masquerading under the rhetoric of protecting children and preventing sex trafficking that will be analyzed in much greater detail in chapters 3 and 4.[168]

Facebook's most sweeping standard is its restriction of content displaying "adult nudity and sexual activity," which it considers as "objectionable content." Policies for regulating nudity as obscenity have always been difficult to formalize, and thus the company's definitions here more frequently come as lists of examples of censorable images. The current list of things you should not post images of on Facebook is as follows:

- Real nude adults, where nudity is defined as
  - visible genitalia except in the context of birth giving and after-birth moments or health-related situations (for example, gender confirmation surgery, examination for cancer or disease prevention/assessment);
  - visible anus and/or fully nude close-ups of buttocks unless photoshopped on a public figure; or
  - uncovered female nipples except in the context of breastfeeding, birth giving, and after-birth moments; health-related situations (for example, postmastectomy, breast cancer awareness, or gender confirmation surgery); or an act of protest.
- Sexual activity, including
  - sexual intercourse;
  - explicit sexual intercourse, defined as mouth or genitals entering or in contact with another person's genitals or anus, where at least one person's genitals are nude;
  - implied sexual intercourse, defined as mouth or genitals entering or in contact with another person's genitals or anus, even when the contact is not directly visible, except in cases of a sexual health context, advertisements, and recognized fictional images or with indicators of fiction; or
  - implied stimulation of genitalia/anus, defined as stimulating genitalia/anus or inserting objects into genitalia/anus, even when the activity is not directly visible, except in cases of sexual health context, advertisements, and recognized fictional images or with indicators of fiction.
- Other sexual activities including (but not limited to)
  - erections;
  - presence of by-products of sexual activity;
  - stimulating genitals or anus, even if above or under clothing;
  - use of sex toys, even if above or under clothing;
  - stimulation of naked human nipples; or
  - squeezing female breast except in breastfeeding context.
- Fetish content that involves
  - acts that are likely to lead to the death of a person or animal,
  - dismemberment,
  - cannibalism, or
  - feces, urine, spit, snot, menstruation, or vomit.[169]

They note that images are allowed that would otherwise violate these standards if the sexual activity is not directly visible or is not sufficiently detailed or if the image was posted in a satirical, humorous, educational, or scientific context. As we'll see below, these standards contain a number of heteronormative and Western biases that become apparent when they are put into practice. It is worth noting first that Facebook leverages the rhetoric of protecting children and the feminist discourse against revenge porn to justify their overbroad implementation of these standards. In short, the company defaults to removing any and all sexual imagery on the platform unless there is a specific "carve-out" that has been codified to allow that content.[170]

Each carve-out that the company has worked into its Community Standards and its "human algorithms" for reviewing flagged content is the result of a previous failure of the system and usually has only been implemented after strong grassroots organizing among Facebook users who have faced multiple account bans and had many pieces of content censored. Take, for example, the carve-out for images of breastfeeding, a hard-won victory by grassroots organizers well documented by Tarleton Gillespie in his book *Custodians of the Internet*. For more than a decade, Facebook censored all images of breastfeeding. As Gillespie notes, "Some women spoke of feeling ashamed and humiliated that their photos, and their experiences as new mothers, were being judged obscene. Removal of a photo was literally an erasure of the woman and her accomplishment, and could easily feel like a particularly personal violation of a deeply held conviction."[171] This censorship also synced up with people's experiences with the hypersexualization of female-presenting breasts that made breastfeeding in public a social taboo, extending these same restrictions into the digital world. In this example, we can see Facebook kowtowing to their advertisers, protecting their brand image, and appeasing heteronormative conservatives at the expense of breastfeeding parents. The policy was not changed until 2014, after a decade of parents politically organizing, engaging in political action, making negative headlines for the company, and engaging in acts of "platform disobedience," like purposefully posting images that violated this community standard and then documenting and sharing Facebook's responses—many people had their accounts banned multiple times to achieve this victory.

More recently, Facebook has been struggling to figure out how best to handle what it refers to as "cultural nudity," which is most often exemplified by photos of "aboriginal" peoples and religious figures and rituals. For

the people in these images, nudity is a cultural norm, and it would be difficult to impossible to visually represent them on the platform without violating Facebook's Community Standards. In 2019, the company convened four cross-functional working groups, engaged in seventeen conversations with external stakeholders, and made a presentation at their Product Policy Forum without reaching a consensus on how to address this gap in their current policies. Facebook's current policy would be to ban any of these images that contain visible female-presenting nipples or the genitalia of either sex, but the company has considered allowing images of age-appropriate people whose female-presenting nipples are exposed, provided the images are not sexually suggestive and no genitalia are visible. This proposal produces some problems, though, as it leads to value judgments of content, the problem of having moderators attempt to interpret content, and would require a lot of labor to overhaul the automated content moderation systems. The other proposed change would add carve-outs for full nudity when there are indicators that the nudity is "cultural/Indigenous" or is posted in the explicit context of "pregnancy/motherhood." This latter proposal presented problems of determining the consent of the nude people depicted; produced the possibility for false positives within the carve-outs; requires a definition of cultural nudity, which is difficult to produce; may produce a slippery slope in which context is used to justify more and more nudity on the platform; and requires an overhaul of the automated system. In the end, the company failed to reach a consensus on new changes and continued to analyze its options for the rest of 2019 while the ban on depictions of genitalia or female-presenting nipples remained the status quo.[172]

From this example, we can see a number of problems in the way content gets moderated on Facebook. First, the company's quest to produce a universal—and English-language-based—set of standards runs into problems when applied in practice. Here these problems consist of cultural variations in what is considered "nudity" and in which contexts that nudity becomes "obscene." While it is easy to see how a company with the practicality of an entrenched hacker ethic would lean on the visibility of female-presenting nipples as the sole determinant of whether an image is obscene, this comes at the expense of hardcoding heteronormativity into the platform. It reinforces the sexualization of female-presenting bodies and produces an unfair double standard that is disadvantageous for female-presenting people.[173] Further, even the proposed policy changes would not introduce toplessness for

female-presenting people and potentially achieve the hoped-for desexualization of female-presenting bodies among populations that have historically suppressed this behavior. As the company noted in its Product Policy Forum, "We are trying to focus on historic, social and cultural norms that exist across different groups of people, whether that's religious groups, racial groups, social groups."[174] This move is also grounded on a homogeneity across historic, social, or cultural opinions about female-presenting toplessness that does not exist. For example, in the United States, thirty-seven states do not have official policies on female-presenting toplessness, and it is more often legislated at the local level, producing a wide variety of local ordinances governing the exposure of female-presenting breasts in the United States.[175]

The most heteronormative aspect of this particular community standard, however, is the assumption that "female nipples" correlate with genitalia or that anatomical sex can be inferred in cases where genitalia are obscured from the shape and size of the breast tissue that visible nipples are attached to. Take, for example, Courtney Demone's #DoIHaveBoobsNow? Project, in which Demone, a transgender woman, posted photos of her exposed chest as she underwent hormone replacement therapy, challenging Facebook and Instagram to answer her very simple question, "At what point in my breast development do I need to start covering my nipples?"[176] Platforms cannot answer this question based on their cisnormative community standards. In other words, their policies turn out to be profoundly *im*practical for regulating the visibility of gender fluid, nonbinary, and trans bodies on the platform. With the current accuracy rates of computer vision-based automated content filtering, coupled with the "human algorithms" Facebook employs to sort out the content that slips through its automated filters, the company can quite accurately identify all images with exposed nipples. So why not just produce an overlaid filter that blurs an image until people confirm that they are willing to be exposed to it, as Facebook already does with violent and medical images posted to the site? Why not simply produce a nipples/no nipples toggle in a user's Facebook settings? It would be easy to implement such a solution and, since the company already verifies user identities and ages, it would be easy to gatekeep children from "unwanted exposure." Heteronormativity on a platform like Facebook is essentially this: to see biased filtering as the default, most practical solution to the problem of content moderation rather than recognizing the ease with which less normative filtering could be achieved across the platform.

Another recent point of contention in Facebook's Product Policy Forum has been how to moderate sexual activity in art. In late 2018 and early 2019, the company convened two working group meetings, consulted with fourteen external stakeholders, and held a presentation at their Product Policy Forum to reaffirm their status quo and formalize better operational guidelines to ensure more consistent enforcement by their "human algorithms." As the company notes, "Our policies distinguish between real world and digital art in the context of adult nudity and sexual activity because we have historically found that digital images are hypersexualized."[177] In essence, the company maintains a medium-specific set of standards. If your image is a photograph of an oil painting or a sculpture, Facebook assumes it is less likely to be hypersexualized than a digitally produced image. The company also is more likely to remove performance art and mass-produced, video-based content. Thus, content moderators are trained to look at whether an image is of a "real object" (paper, wood, canvas, wall) or composed in a "traditional medium" (water colors, pencil, marker, charcoal, spray, marble, bronze) and to use Google Image Search to check whether the art is "real or digital." They also look to see if an image has been altered by photo editing software through things like cut and paste signals and whether there are traces of paint programs, like vector shapes or 8-bit lines. This essentially reproduces a traditional classist distinction between "high" and "low" art in which Tom of Finland erotica is permissible because it's been made with oil paints on canvas but anonymously created digital nude portraits are not. While it might not be readily apparent how this connects to heteronormativity, one needs only remember the financially precarious existence of much of the LGBTQIA+ community, particularly a large section of the trans community. This makes it more difficult for the community to publicize its arts and culture, arbitrate and litigate censorship, and take in essential revenue to maintain and produce new artworks. As smartphones increasingly become a necessity for everyday life—essential for gaining employment, managing finances, obtaining housing, and so on—they are increasingly the windows through which arts and culture are accessed, disseminated in the public sphere, and marketed to generate revenue. Digital communications mediated by these smartphones and the social media platforms that dominate their internet usage can easily lead to a future in which artistic production is less viable for the LGBTQIA+ community. By producing a medium-specific set of community standards surrounding artistic nudity, Facebook is potentially leaning into a future in

which LGBTQIA+ art bears an undue burden of censorship and risks being rendered invisible to the broader public.

Taken collectively, despite response to public pressure to improve their content moderation policies and practices, the operation of Facebook's "human algorithms" leaves much to be desired. The employees responsible for the moderation of content of central importance to LGBTQIA+ communities in the United States are largely the least experienced, operate from other cultural contexts, and are underpaid and overburdened, leading to uninformed, exceedingly fast decisions about censorship. As I've shown, these decisions are particularly impactful when it comes to content surrounding sex work, breastfeeding, nipple exposure more broadly, cultural nudity, and artistic expression. This creates an undue burden of censorship on LGBTQIA+ communities and disallows a large portion of content that many people would not find to be explicitly obscene or pornographic, ranging from community building, social activism and organizing, sex education, explorations of gendered embodiment, and artistic expression. As we'll see in chapter 3, these actions by Facebook are not the exception but the rule on the internet. LGBTQIA+ content is faced with undue censorship, account banning, and demonetization across the web.

The Digital Closet

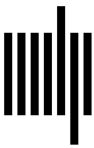
How the Internet Became Straight

By: Alexander Monea

Citation:

The Digital Closet: How the Internet Became Straight
By: Alexander Monea
DOI: 10.7551/mitpress/12551.001.0001
ISBN (electronic): 9780262369138
Publisher: The MIT Press
Published: 2023

The open access edition of this book was made possible by generous funding and support from MIT Libraries, and MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Bembo Book MT Pro by the MIT Press.

Library of Congress Cataloging-in-Publication Data

Names: Monea, Alexander, author.
Title: The digital closet : how the internet became straight / Alexander Monea ; foreword by Violet Blue.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Series: Strong ideas series | Includes bibliographical references and index.
Identifiers: LCCN 2021019772 | ISBN 9780262046770 (hardcover)
Subjects: LCSH: Internet—Social aspects. | Homophobia. | Sexism.
Classification: LCC HM851 .M6593 2022 | DDC 302.23/1—dc23
LC record available at https://lccn.loc.gov/2021019772

10 9 8 7 6 5 4 3 2 1