(**Pre-published version)** From: Margaret Gallagher and Aimée Vega Montiel (Eds), 2023. *The Handbook of Gender, Communication, and Women's Human Rights.* Wiley-Blackwell

# Power in AI: Inequality Within and Without the Algorithm

**Kate Devlin**
**King's College London**

## Introduction

Artificial intelligence (AI) lacks a standardized definition but, in general, is the concept of machines being able to carry out tasks in an intelligent manner. The degree and definition of intelligence is up for discussion, as is the autonomy of the machines. That said, AI exists in the world today in reasonably narrow and specific contexts, generally in the form of software that automates processes by handling large amounts of data and deriving patterns within it.

Most of the applications of AI we encounter are from a subset known as machine learning: the computer learns from data we feed it, analyzing patterns and looking for connections. Sometimes the data are clearly labeled and the software can check for similarities, classifying information at far greater speeds than a human. In other instances, the machine looks for commonalities and clusters of features in unlabeled data and discovers new and interesting patterns. At no point does the AI *understand*; it merely works out rules, either pre-defined or self-imposed.

This chapter explores AI's gendered history from its birth as a discipline through to the pre- sent, giving an account of the major events in the field set against the wider techno-social landscape of the time. It explores the state of play today, where power is corporate and techno- cratic, and bias is not only a factor in the world of employment but embedded in the very algo- rithms themselves.

## The Conception and the Birth of AI

The idea of thinking machines has a long history; tales of automata appear over the centuries in myths from around the world (Cave et al., 2020). Many exist only in stories, such as the Greek Homeric tales of bronze servants in the hall of the gods, but these narratives persist down the years (Liveley & Thomas, 2020). Elaborate medieval tales of mimetic creatures and clockwork mechanicals survive today (Truitt, 2020). By the eighteenth century, skilled inventors were pro- ducing sophisticated figurines that had the appearance of life-like capabilities. These automata gave only the illusion of intelligence but the concept of creating machines that could perform self-directed labor was a topic of interest, particularly with the arrival of the Industrial Revolution.

In 1822, the English mathematician and engineer Charles Babbage began to design his Difference Engine to mechanize calculations. It was never finished but it inspired his next and more complex model, the Analytical Engine. That was also incomplete on his death, but Babbage had made a theo- retical leap, envisaging a mechanical calculator based on the operation of a Jacquard weaving loom. Fabric with a Jacquard weave has a notably complex pattern, as seen in brocades and damasks, and was created using punched cards holding instructions in a binary system. The arrangements of punched holes signified where the hook holding the thread should rise and fall. Babbage realized that mathematical instructions could be passed the same way (Morrison & Morrison, 1961).

Prior to the Industrial Revolution, weaving was carried out in people's homes. It was predom- inately viewed as women's work as it was women who spun the yarn (Burnette, 2008). With the rise of factories, the gender division in the textile industry changed and men took over some of the more skilled tasks but, as women were paid around one-third of the rate that men received, it was still cheaper to hire a female labor force. Such a shift was later seen in computing where women's labor programming early machines was side-lined when personal computers (PCs) became popular and the role was re-evaluated to recognize the skill involved – and deemed better suited to a man (Ensmenger, 2015; Hicks, 2017).

In 1833, Ada Countess of Lovelace accompanied her mother to visit Babbage's Difference Engine. Already interested in mathematics, she collaborated with Babbage on the Analytical Engine, which she described as "an embodying of the science of operations." Her vision was "that the Analytical Engine weaves Algebraical patterns, just as the Jacquard loom weaves flowers and leaves" (Morrison & Morrison, 1961, p. 252; Plant, 1995, p. 50).

Today, Lovelace is an inspirational figure for women in technology and is widely considered the first computer programmer for her insight into how computing could be applied to any pro- cess. She was interested in the workings of the brain and keen to model "a calculus of the ner- vous system" (Woolley, 2002, p. 305). However, she was clear that machine cognition was not within the remit of the Analytical Engine, stating "it can follow analysis; but it has no power of anticipating any analytical relations or truths" (Lovelace, 1843, p. 722).

Almost a century after Lovelace's death, the British mathematician Alan Turing published the paper "Computing Machinery and Intelligence" (Turing, 1950). Computer science was still a new field: the world's first electronic digital programmable computer had been built in 1944 at Bletchley Park, where Turing worked as a wartime codebreaker. His 1950 paper is a seminal work in the advent of AI. He opened with the question: "Can machines think?" and went on to debate the prov- ocation by attempting to refute arguments against his idea of machine intelligence. He addressed Lovelace's claim that machines could never do anything original. Turing believed machines could be surprising, even if this was confined to working out unforeseen consequences from data. Today, machine learning can do exactly that. Turing's hounding and subsequent death in 1954 (he was prosecuted for homosexual acts – then illegal in England – and died by suicide) meant he did not live long enough to see the formation of a dedicated field of research two years later.

The Dartmouth Summer Research Project on Artificial Intelligence in New Hampshire, USA, accepted as the founding of the field, was conceived by four researchers in 1956: John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. Their proposal was that "a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956" (McCarthy et al., 1955, para. 1). The event was ground-breaking, setting the direction of the field for decades to come, and has been described as the birth of AI (Russell & Norvig, 2021). In the end, it was attended by 20 researchers – all of them white men (Solmonoff, 1956). Minksy's wife, Gloria, was behind the camera to record the group for posterity (Bell, 2021).

That no women actively participated in the Dartmouth workshop (aside from an unnamed secretary, gender unknown, mentioned in the proposal) is not surprising. At the time, women's involvement in computing was generally in the low-status work of the coder, issuing instructions to the machine (Ensmenger, 2015; Hicks, 2017). These complex tasks were neither recognized nor feted, despite their contribution to advancements in areas such as cryptography and aerodynamics. In-group hierarchy existed too: women were segregated by race, with Black women treated more unfairly than their white counterparts (Shetterly, 2017). Women were also broadly excluded from the professoriate. Rossiter describes how, in the mid-1950s, "the largest proportions of women faculty were at the poorest and least prestigious colleges and universities" (1995, p. 187) and that research grants were often closed to them. The birth of the new discipline of AI was a birth with no mother, and the founding fathers of AI were creating machines in their own image.

The Dartmouth workshop heralded an age of optimism: researchers declared there would be breakthroughs "in 10 years" or "within a generation" – timescales that inevitably shifted as dead- lines approached. For nearly two decades, things seemed excitingly possible and advancements in areas like search algorithms and speech recognition apparently supported this. However, the baseline was low and so all leaps seemed huge.

By the 1970s, the high expectations of the initial research had fallen short. The complexity of the discipline became apparent. AI research faced strong criticism (Lighthill, 1973). The money, which had flowed easily since Dartmouth, dried up. Progress halted, and 1974 marked the beginning of the first "AI winter," where pessimism replaced the original hype.

## The Rise of the Personal Computer and the Fall of Women in Computing

It took nearly another decade for AI funding to pick up again. In the 1980s, the rise of the "expert system" (one that uses logical if-then-else rules to mimic human expert decision making), and a divergence into knowledge-based systems, meant research progressed once more. It was during this period that PCs began to dominate the market. This solidified the demographic shift from computing as rote and repetitive feminized work to male hacker culture. Women were pushed out by professionalization and masculinization (Ensmenger, 2015). The number of women studying computer science peaked in 1984, then began to decline, never fully recovering. Wendy Hall and Gillian Lovegrove, writing about women in AI in the 1980s, remarked:

*It is quite feasible to argue therefore, that the low number of women in computer science will result in a correspondingly low number of women in AI. If this is the case, the effect on AI research could be more profound than on computer science research in general. This is because the kind of qualities that are important to the AI community in particular, are often found in women: for example, attention to detail, organisational skills and the ability to communicate easily and effectively.* (1988, p. 270)

By the late 1980s, another AI winter had set in, this time a combination of unmet expectations and a new market emphasis on desktop computing (Anyoha, 2017). Funding picked up again in the 1990s as advances were made across many areas of AI. There were notable contributions by women, particularly in robotics: Cindy Mason and Carol Stoker at NASA, and Leslie Pack Kaelbling, Lynn Andrea Stein, and Cynthia Breazeal at MIT (Massachusetts Institute of Technology). Nonetheless, the iconic AI moment of that decade was IBM's computer, Deep Blue, beating world-champion chess player Garry Kasparov – chess being viewed as the pinnacle of intellectual activity and, erroneously, as a signifier of innate male superiority (Bilalić et al., 2009). Ironically, the success of Deep Blue was not due to any sudden leap in AI but rather to an improvement in computer capabilities that brought ideas from decades before into fruition by enabling exhaustive searches of possible moves.

## Deep Learning

When the new millennium arrived, the hopes of the 1950s had still not been realized. However, the increase in computer capabilities, combined with the proliferation of data and the resurgence in machine learning techniques, meant AI began to offer practical and useful applications. Much of this was due to the advancement of artificial neural networks.

Most animal brains have neurons. Individually, neurons transmit nerve impulses, taking input (stimuli from sensory organs) and passing it through networks of neurons, processing the information. The output is a response by the nervous system. *Artificial neural networks* are a software simulation loosely modeled on a biological brain. In simplified terms, they are com- prised of an input layer, a hidden layer, and an output layer, all made up of interconnected processing nodes. Input values fed in to the first layer are processed in the hidden layer where calculations are carried out and decisions are made, then these values are output. For example, if the network was given a picture of a cat as input, the hidden layer would carry out calcula- tions to determine whether the picture was of a cat and, if it was, it would output "cat." If the hidden layer calculated that the picture did not meet its threshold for resembling a cat, it would output "not cat."

Artificial neural networks had been proposed in 1943 when neurophysiologist Warren McCulloch and mathematician Walter Pitts published a paper describing neurons and modeled them using an electrical circuit (McCulloch & Pitts, 1943). This was followed up in 1957 by Frank Rosenblatt's Perceptron model (Rosenblatt, 1957), the first trainable version, broadly meaning that it could be taught to fit the objectives given to it. Further advancements in artificial neural networks occurred in the late 1970s and again in the mid-1980s, but it took increased computing power – specifically the graphics processing unit (GPU) – and the vast increase in user-generated data for it to be realized as *deep learning*.

The term "deep learning" was coined in 1986 by a woman, Rina Dechter (Dechter, 1986). However, the phrase was not used to refer to artificial neural networks until 2000. Deep learning as we know it today is a subset of machine learning. The word "deep" refers to the number of layers a neural network has: a simple artificial

neural network is limited in the amount of learning it can do but if many layers of processing are added, the performance increases drastically (Kelleher, 2019).

An important contribution to deep learning was also made by a woman: Fei-Fei Li founded a large visual database, ImageNet,[1] after realizing that machine learning algorithms were limited by lack of data. How could algorithms recognize the contents of an image, for example, if they only had a few images to work with? In 2006, she approached another woman, Christiane Fellbaum, a co-developer of WordNet, a lexical database used as the basis for automatic text analysis. Li started to build the dataset and, much like weaving before the factories, outsourced the labeling to home workers via Amazon's Mechanical Turk (Gershgorn, 2017). In 2009, her team published their ImageNet paper (Deng et al., 2009). In it, they remark that Amazon's Mechanical Turk was particularly suitable for labeling due to a global user base, but a study at that time showed 57% of "Turkers" were located in the US, 32% in India, and the remaining 11% split across a number of countries, the more sizeable share of which were Western (Ross et al., 2009). By 2012, ImageNet had become a benchmark for image classification machine learning algorithms and thus played a fundamental role in deep learning becoming state of the art (Marcus, 2018).

## The Field Today

Today, the vast availability of data and the computing power means machine learning is embedded in many of the applications we take for granted. The power to analyze and compare text and images means it is used to classify information, make recommendations, detect anomalies, and provide solutions. Millions – billions – of people encounter machine learning everyday, often without noticing. It is used in voice assistants, mapping software, surveillance, product recom- mendations, medical diagnoses, and financial systems. Its ubiquity is subtle. There were around six billion subscriptions to mobile broadband worldwide in late 2021 (Ericsson, 2021); newer smartphones have their hardware optimized for AI. Each time we use our phone to map direc- tions or take a photo, we're using its machine learning capabilities.

For machine learning to be applied to a task, the data need to be prepared. A source for the data is chosen, a dataset is selected, the data will usually need to be "cleaned" to ensure the machine learning algorithm can process it, and the data may need be labeled via human input. This labeling is required where *supervised machine learning* is used; that is, the algorithm has a description of the data it is given and can classify new data using that baseline. It is supervised in that it has been shown what to do. An example of supervised learning includes labeling (or "seg- menting") images for use by self-driving cars. Before the algorithm can get to work, human oper- ators annotate images of real-world scenes that the vehicle's deep learning algorithms use. This might involve identifying parts of the image such as traffic signs, road barriers, or other vehicles. Similarly, medical images can be labeled to show anomalies like tumors. Machine learning can then look for and identify similar anomalies in new images.

By contrast, *unsupervised learning* allows the machine to look for unlabeled similarities and patterns and thus form its own classifications. This is powerful because algorithms analyzing and comparing hundreds of thousands of datapoints often uncover correlations humans would miss. Recommender systems are a form of this: an algorithm can look for patterns in the TV you watch and suggest other programs by grouping viewers with similar watch histories. Likewise, credit card fraud detection can flag spending patterns that are unusual and do not fit a user's reg- ular spending profile.

Another notable form of machine learning is *reinforcement learning*, which takes the form of a trial-and-error approach combined with a reward system: when the algorithm must make a decision, it tries multiple solutions. When a solution succeeds, this is noted as a reward. If it fails, it can be discounted. This way, the algorithm can improve itself. It can be used in areas where feedback improves performance: refining recommender systems, for example, where success can be measured by end-user engagement.

In all these forms of machine learning, bias occurs within the algorithm – from selecting data to processing it through layers of neural networks (Ntoutsi et al., 2020). It also occurs outside of the algorithm too. The environment in which software is developed involves choices where bias can influence the output even before the data is selected.

## Within and Without: The Algorithm and the Environment

An examination of two of the main uses of machine learning – processing language and classi- fying images – exposes how bias and discrimination can occur algorithmically. Natural language processing (NLP) is a subfield of machine learning where human communica- tion via language – speech and written – is analyzed computationally to read in and/or to output human-like text. This is no trivial matter. Human conversation is nuanced, contextual, and ambiguous. For a machine to process this as if it "understands" what is being said requires the input to be broken down into pieces, determining the relationships between words and phrasing. To capture those semantic relationships, a technique known as *word embedding* can be used, where words are weighted numerically so the likelihood they will co-occur can be identified. This gives the machine a mathematical representation of context.

Although hundreds of papers have been written on NLP since its first iteration in the 1960s, and thousands of systems deploy the technique, only in the last decade has the problem of bias – particularly gender stereotyping – been explicitly addressed (Bolukbasi et al., 2016; Hovy & Prabhumoye, 2021; Schmidt, 2015). Linguists were aware of this – it is a well-researched phenomenon – but as the results of algorithmic discrimination became more apparent, greater scrutiny was directed toward computational processes often taken for granted; indeed, often taken for granted because of a desire to see the computer as an arbiter of neutrality, distanced from human flaws.

Bolukbasi et al. demonstrated how NLP word embedding algorithms, trained on a dataset of Google News stories, associated words such as "homemaker," "nurse," and "guidance counselor" with women, and "captain," "financier," and "boss" with men. "In other words," they write, "the same system that solved the above reasonable analogies will offensively answer 'man is to com- puter programmer as woman is to x' with x=homemaker. Similarly, it outputs that a father is to a doctor as a mother is to a nurse" (2016, p. 3). This echoes a well-known issue with Google's translation software that converted Turkish sentences with genderless pronouns "O bir doktor. O bir hemş¸ire" to the English sentences: "He is a doctor. She is a nurse." This arose in part because "Google Translate learns from hundreds of millions of already-translated examples from the web" according to Google Translate's Product Manager (Kuczmarski, 2018, para. 2).

One headline example of gender bias in NLP occurred in 2015 when Amazon automated their recruitment process using machine learning. Amazon used NLP to score submitted resumés against those of successful candidates from the previous decade. The machine learning algorithm determined linguistic patterns from successful resumés but in doing so reflected human hiring bias that had previously been in play. Because women were underrepresented in the historical data, men's resumés were signal boosted (Righetti et al., 2019). Resumés were penalized for including the word "women" – such as "women's chess club captain." The algorithm taught itself that the ideal candidate was someone with a resumé that read as "male." Existing bias had created a feedback loop (Gebru, 2020). Amazon abandoned the trial.

Gender is not the only bias seen in word embeddings. Caliskan et al. (2017) also identified racial bias. They replicated earlier non-AI studies on bias using an off-the-shelf, widely deployed machine-learning model and found the word embeddings favored European American names (e.g. Brad, Brendan, Geoffrey, Anne, Carrie) over African American names (e.g. Jermaine, Jamal, Leroy, Latoya, Tanisha). Likewise, when repeating the study with resumés, those with European American names were more likely to gain an interview than those with African American names. This shows "that if we build an intelligent system that learns enough about the properties of lan- guage to be able to understand and produce it, in the process it will also acquire historical cultural associations, some of which can be objectionable" (Caliskan et al., 2017, p. 186).

Work by Hutchinson et al. demonstrated that widely used machine-learned models classed words associated with disability as more negative. They found, for example, the statement "I am a person with mental illness" scored as being highly "toxic" in comparison to the statement "I am a tall person," which scored as neutral (2020, p. 1). The algorithm sets and perpetuates an expectation of a normative standard. This can be seen in everyday interactions with AI such as completing captchas to gain access to a website, or issuing a voice command. This is non-trivial for people with disabilities where their speech is impaired or their coordination is slow. As AI Now's report on Disability, Bias, and AI puts it: "In all of these cases, in order to proceed with a desired action, you need to prove to the machine that you are a certain type of human" (Whittaker et al.,

2019, p. 14). Such cases sit alongside higher-level examples of ableism – a prevailing benev- olent paternalism where technological "cures" are favored, such as viewing conditions like ASD (Autism Spectrum Disorder) as a medical deficit requiring solutions through software and robots to improve integration with "normal" society.

Hovy and Prabhumoye (2021) broadened the critique of NLP further: age is a factor too, as is language and dialect. Most NLP software focuses on English as the default language. English has the most commercial demand, so it perpetuates development in that area due to avail- ability, despite a desire and a need for multi-lingual work. As Steven Bird points out, "language loss is often the by-product of oppression" (2020, p. 3507). Throughout history, indigenous languages have been erased as a form of cultural imperialism. Language is a marker of identity, and the defaulting of NLP to English entrenches digital colonialism. And, adds Bird, "for the remaining 90% [of the world's languages], language tends to be oral, emergent, untranslatable, tightly coupled to a place." He reminds us that while the dominant Western approach to language may have "a sense of entitlement: to protect, to save, to know, to mine," there is a far greater unacknowledged role of indigenous languages as intrinsically tied to intangible culture, and to be truly diverse means centering indigenous voices and community aspirations. That requires rethinking machine learning's current view of language as modularized data that can be represented mathematically.

Bones et al. highlight that language matters not just in terms of the working of algorithms but also in the way AI itself is described in terminology, jargon, and conversation. The way in which we communicate our concepts shapes our perception. The term "data" is conceived as "something that exists independent of human creation" – it has connotations of something to be collected, stored, or reported. Echoing Bird's warning, this in turn distances us from the notion that data is personal, instead portraying it as capital. By contrast, computer hardware and software are humanized with anthropomorphic terms such as "memory" and "learning," giving agency and independence to the machine while also feeding into narratives of fear around superintelligence (Bones et al., 2021, p. 31).

## Algorithms: Images

Deep learning has had a profound effect on image classification. Where once this task was an intractable goal of computer vision research, now AI can be used to sort, analyze, and identify photographs and videos at lightning speed. A computer cannot semantically understand an image. It receives a mathematical representation of that image in the form of arrays of pixel data. If we consider how humans view the world, we learn what objects are (e.g. a bicycle) and we can identify that object even when it is moving, when it is lying on the ground, when it is propped against a wall, or when we view it from different angles. A computer has no such knowledge. It can only identify a bicycle if someone has labeled enough images so that stated features can be spotted, or it has seen enough unlabeled images that it can spot recurring features and therefore develop its own internal representation of "bicycle." When machine learning classifies an image, it is actually assigning a probability that the image it has analyzed has enough features to meet a description threshold.

In 2012, a team from the University of Toronto trained a deep convolutional neural network, AlexNet, to classify the 1.2 million high-resolution images in the ImageNet large-scale visual rec- ognition challenge. It won by a large margin and the accompanying academic paper became one of the most influential papers ever published in computer vision (Krizhevsky et al., 2012). In the same year, Google deployed face-detecting algorithms on YouTube and set them to work, without supervision, to see what they could learn from ten million unlabeled cat videos (Le et al., 2013). It showed a computer could be given huge amounts of unlabeled data and could pick out a pattern we humans know as "cat" without ever having been told what a cat looked like in the first place.

This deep learning revolution of 2012 meant that image classification could be rolled out into a variety of applications: analyzing medical imagery, identifying objects in satellite pictures, orga- nizing photographs on social media, and enabling vision for autonomous vehicles, for example. One particularly contentious area that benefitted was facial recognition. Although it had been trialed in the 1960s and had some commercial success from the 1990s, deep learning enabled it to flourish.

Facial recognition has four main components. The first step is facial extraction, where the algorithm identifies the area of the image corresponding to the face and isolates it. Next, the extracted face undergoes

normalization, e.g., correcting lighting to make it more even, adjusting for angle, etc. Certain comparative features and measurements are then extracted and checked against a database to look for matches. For example, Facebook's DeepFace uses three-dimensional data points to extract features, combined with deep learning trained on four million photographs uploaded by Facebook users (Taigman et al., 2014).

The rapid uptake of facial recognition led to increased awareness that it was fundamentally flawed. Klare et al.'s study in 2012 found the algorithms consistently had lower matching accu- racies on faces that were female, Black, and aged 18–30 (Klare et al., 2012). It was not the first paper to evidence shortcomings in the process, but it was one of the first to deal with it in machine learning. They advised that using more demographically diverse training data could alleviate much of the problem. A 2016 article in *The Atlantic* flagged that people of color were more likely to be overrepresented in mugshot and surveillance databases due to policing bias. "In other words," they wrote, "not only are African Americans more likely to be misidentified by a facial recognition system, they're also more likely to be enrolled in those systems and be subject to their processing" (Garvie & Frankle, 2016).

A decade on from Klare et al.'s study the situation has still not been rectified. Despite more and more being revealed about the dangerous bias in these systems, there have still been egre- gious errors of judgement. When the American Civil Liberties Union (ACLU) tested Amazon's facial recognition system in 2018, the software incorrectly matched 28 members of Congress with other people's arrest photos, disproportionately targeting people of color (Snow, 2018). At this stage, Amazon's technology was already being used by police agencies. Crawford, describing mugshot databases, states: "Machine learning systems are trained on images like these every day – images that were taken from the internet or from state institutions without context and without consent. They are anything but neutral" (2021, p. 94).

In 2015, Google publicly apologized when its Photos software automatically classified a Black couple as gorillas. By 2017, however, it had only implemented a workaround: when *WIRED* mag- azine tested the software, it found Google had simply removed the search terms "gorilla," "chimp," "chimpanzee," and "monkey." Google confirmed it had censored those words from searches and said it was working on a longer-term fix (Simonite, 2018a). In 2019, the *New York Daily News* revealed Google was refining its facial recognition software by testing it on homeless people of color in exchange for $5 gift cards. The newspaper reported an ex-Google staffer as saying "They [Google] said to target homeless people because they're the least likely to say anything to the media...The homeless people didn't know what was going on at all" (Otis & Dillon, 2019). Google defended the study but said it was investigating the claims around participant consent.

Outside of Silicon Valley, facial recognition is being used to actively discriminate. Although technology-driven surveillance in China is widely deployed for general social monitoring, it is ramped up in Uyghur areas (Leibold, 2020). Chinese corporations and academics have published articles that claim to distinguish Uyghur people via facial recognition alone (Daly, 2019).

Skin color is not the only characteristic dangerously misidentified by the image processing algorithms. Jutta Treviranus, a Professor of Design, writes about trying out machine learning models to guide automated vehicles through intersections. She gave the models footage of her friend who uses a wheelchair:

*When I presented a capture of my friend to the learning models, they all chose to run her over. I was told to try again once the learning models had been exposed to more data about people using wheelchairs in intersections. I was told that the learning models were immature models that were not yet smart enough to recognize people in wheelchairs, they would expose them to learning data that included many people using wheelchairs in inter- sections. When I came back to test out the smarter models they ran her over with greater confidence.* (2018)

As Whittaker et al. remark, "As the use of these systems expands into an increasing array of sensitive social domains, the risk of not being 'seen' as human become greater. As with the exam- ples of autonomous vehicles, if someone doesn't 'look like a pedestrian' when crossing the street, they risk being killed" (2019, p. 15).

Image recognition has been shown to replicate and amplify biases in the real world (Buolamwini & Gebru, 2018; Schwemmer et al., 2020). Much like the NLP word embeddings, the labeling and classifying of images of women matches the inequalities and gender stereotypes we see in society. A critique of ImageNet by Crawford and Paglen explores the problematic categorization of the images, exposing stereotyping and offensive labeling, pointing out that "these training sets have been down- loaded countless times, and have made their way into many production AI systems and academic papers" (2021, Missing Persons section, para. 6). Safiya Noble's research has shown how gender, race, and class intersect: "Representations of women (particularly Black women who are codified as 'girls') are frequently ranked on a search engine page in ways that underscore their lack of status in society" (2013, para. 5, 2018). The perpetuation of this racism and misogyny means algorithmic injustice is entrenched in software used daily worldwide. The feedback loop strikes again.

When systems do categorize gender, they categorize it in binary terms. Costanza-Chock writes an account of being (mis)classified by airport body scanners because "at each stage of this interaction, airport security technology, databases, algorithms, risk assessment, and practices are all designed based on the assumption that there are only two genders, and that gender presenta- tion will conform with so-called biological sex...Most cisgender people are unaware of the fact that the millimeter wave scanners operate according to a binary and cis-normative gender con- struct; most trans people know, because it directly affects our lives" (2020). ImageNet is no exception: in its taxonomy images are sorted as "male body," "person," "juvenile body," "adult body," and "female body" (Crawford & Paglen, 2021).

Another concerning development is colloquially termed "AI phrenology." Phrenology, and the closely related physiognomy, is a pseudoscience that claims to be able to determine character traits via appearance, such as the idea that criminals have large jaws or sloping foreheads. Phrenology and physiognomy were used to justify racial persecution for centuries before falling out of favor in the mid- to late- 1800s. Nonetheless, a spate of proposed and published papers have (erroneously) claimed to be able to use machine learning and image processing to detect such traits as criminality and trustworthiness (see Agüera y Arcas et al., 2017 for a comprehensive debunking).

In 2017, two researchers from Stanford University in the USA claimed they could use machine learning to determine people's sexuality just from a photograph of their face. Coined the "gay- dar" paper (a portmanteau of gay and radar, slang for being able to determine someone's sexu- ality through intuition), the study immediately provoked cynicism. When criticized as being irresponsible, if not downright dangerous, the researchers replied they had wanted to raise the alarm (Murphy, 2017). However, their work could endanger the lives of gay people in countries where homosexuality is punishable by death. The study was later replicated by a Masters student from the University of Pretoria in South Africa (Leuner, 2019) and even when faces were com- pletely blurred out, the software was still able to "predict" sexual orientation, suggesting the algorithm was not analyzing facial structure at all but relying on other cues (Quach, 2019). And yet the potential for weaponization remains.

## Environment

As the "gaydar" debacle shows, even before dataset choices are made bias is already in play: the choice to create a piece of software is subjective and the starting point is rarely impartial. Meta (formerly Facebook), for example, owes its origins to its founder, Mark Zuckerberg, creating FaceMash – a webpage that showed the user photographs of two women and asked them to rate which one was "hotter" (Carlson, 2010; Kaplan, 2003). Zuckerberg has played down this misogyny in recent years (Shaikh, 2019).

The tech sphere is known to be skewed against – and sometimes actively hostile to – women (Klinger & Svensson, 2021). The statistics reflect this: a 2018 World Economic Forum report showed only 22% of AI professionals globally were female, and this was still below 25% by the time their 2021 report was published (World Economic Forum, 2021, p. 60). A *WIRED* and Element AI study in 2017 found only 12% of leading machine learning researchers were female (Simonite, 2018b). In the US, only 18% of women study computer science at undergraduate level; in the UK, it's 16%. The gender gap is pervasive, but it is not a constant: global variations occur. In India, for example, 45% of computer science university students are female (West et al., 2019a). An analysis of global data from 2017 pertaining to ICT professionals showed that the gender gap is greater in the eco- nomically developed regions of Europe and North America. By contrast, African and Asian

Pacific regions were nearly twice as likely to have women in tech roles (Chow & Charles, 2019). Despite this, the US-based origins of AI and the current Western dominance of the discipline mean that gender representation in the field follows the trends in those countries.

The treatment of women in the work environment is typified by high-profile cases over the past decade. In a 2016 Bloomberg article, Margaret Mitchell, the only woman in her Microsoft Research group, described AI research as "a sea of dudes" (Clark, 2016), provoking extensive debate on social media. In October that year she left her role, stating "Microsoft has let me know that I am not making enough impact, and promoted the men around me who helped with the project" (Mitchell, n.d., Timeline section Late August 2016). Mitchell's experience was merely one occurrence in a long line of sexist behavior in the AI world. In 2017, Uber agreed to pay $4.4 million into a fund to compensate employees who experienced sexual harassment, following whistleblowing by one of their female engineers (Sonnemaker, 2019). In 2021, Google settled a class action lawsuit by agreeing to pay $3.8 million to 5,500 of their employees and job applicants to compensate for pay disparities and hiring discrimination (US Dept of Labor, 2021).

AI is technocratic, and today the power resides in Silicon Valley – where the money is. (China, though, is hot on its heels.) Funding plays a significant role in determining what technology suc- ceeds, but women are less likely to seek and receive funding (Serwaah & Shneor, 2021). Major developments in AI occur in the Big Five corporations: Google, Amazon, Meta, Apple, and Microsoft, and in the US research laboratory, OpenAI. Predominantly male leaderships create products unsurprisingly geared toward the same audience as them. When Apple released a health tracker, it didn't include a way to track menstruation (Eveleth, 2014). When voice assistants were first launched, all of them had female voices – a subservient gendered software (West et al., 2019a).

Outside of the "Big Tech" that drives AI, academia is often viewed as a neutral voice that can temper the corporate power grabs. However, it is not immune from the influence of the tech giants: AI researchers are often poached by industry (Ram, 2018), or form collaborations that veer from universities' perceived neutral stance. On occasion, the funding to academic institu- tions can be unethical, such as MIT's decision to accept donations from Jeffrey Epstein following his conviction as a sex offender (Farrow, 2019). Peña & Varon call attention to the "neoliberal consortiums" (2019, p. 31) whereby universities are complicit in the development of systems that are adopted by governments for such uses as predictive policing or community monitoring. With universities being subject to the market, they cannot avoid reflecting the market, and thus the teaching and research itself is subject to commercialization and business-oriented strategies (Troiani & Dutson, 2021) – and with it the systemic biases and discriminations.

Unethical practices in AI employment do not stop at the door of a Silicon Valley office. The labeling and flagging of data required for machine learning has resulted in the largely unregu- lated employment of poorly paid "ghost workers." In 2018, BBC News told the story of Brenda, a woman from the Kibera slum in Nairobi. Brenda's employers were a US company preparing data for AI applications. They justified their outsourcing to Kenyans by claiming they raised people out of poverty, yet commented that "one thing that's critical in our line of work is to not pay wages that would distort local labour markets." They also stressed their pride in a diverse workplace (50% female). However, despite the company's talk of providing opportunities, the news report also flagged that "none of the staff we saw at the office had any kind of accept- able ergonomic support, often crouching over, clicking away furiously, for hours on end – a certain strain to eyes and body" (Lee, 2018).

Likewise, Bartoletti writes about labeling of medical images outsourced to women in India: "It is hard to see a starker contrast than that between the highly paid male execs in Silicon Valley, where the median pay for a top 100 CEO is $15.7 million a year, and the quotidian reality of women in India who are feeding endless images of polyps to the information-hungry machines on which they labour." She goes on to say: "Similarly, many tech workers across the world spend their working days going through violent images to teach a machine what abusive imagery looks like so it can be removed from the internet. This is an unsettling and anxiety-inducing job, with known and unknown psychological consequences for those engaged in it, as employees have reported post-traumatic stress disorder, chronic insomnia and a generalized, debilitating fear. Imagine having to watch scenes of child abuse for hours on end every single day" (Bartoletti, 2018, p. 18). The invisible work takes an intangible toll. Those with the power do not have to face these dull, repetitive, and sometimes traumatic processes.

In their 2019 book *Ghost Work*, Gray and Suri delve into this invisible labor – piecework that, like weaving before the factories, can be carried out at home and in one's own time – necessary but undervalued. They call for us to think seriously and compassionately about the human labor in the loop, to make visible their contributions (Gray & Suri, 2019).

## Toward Dismantling Discrimination

Recognizing that systems in place today are built on inherent and inherited biases, steps toward addressing oppression are required. The field of AI ethics was already aware of this problem, with concern raised about the tangible potential for good and evil dating back to 1948 (Borenstein et al., 2021), but there has been significant growth in the number of papers on the topic from 2016 onward as the discriminatory effects of data and algorithmic bias from deep learning have become more apparent. This includes the proposal of multiple guidelines, principles and frame- works (Dignum, 2019; Floridi et al., 2020; Winfield, 2019) and a wider call for attention to respon- sible research and innovation.[2] Actual regulation, however, is not easy given issues around jurisdiction, corporate pushback, and lack of algorithmic transparency (Hoffmann-Riem, 2020).

A number of companies have decided to look willing by engaging their own ethics researchers, but are not always happy to listen to them. In 2021, Google ousted their lead ethical researcher Timnit Gebru, a Black woman whose job was to study the impact of Google's AI. She was forced out following controversy over her authorship of a paper calling out bias in Google's language software (Simonite, 2021). Margaret Mitchell, who was working alongside Gebru, was fired shortly after. Gebru described the company as "silencing marginalized voices" (her full email is published in Newton, 2020, para. 6). The term "ethics washing" has become a popular way of describing companies that claim to care about the harm their systems might cause while simulta- neously continuing that harmful behavior.

As Gebru's treatment shows, attempts to call out poor practice can be met with hostility, par- ticularly toward women of color. Peňa & Varon propose "a transfeminist critique and framework that offers not only the potential to analyse the damaging effects of AI, but also a proactive under- standing on how to imagine, design and develop an emancipatory AI that undermines consum- erist, misogynist, racist, gender binarial and heteropatriarchal societal norms" (2019, p. 29). Birhane and Guest echo this, calling for a dismantling of the sexist, racist, and colonial roots – and one that must involve more than forming a diversity panel or advisory board populated by white women who often lack intersectionality and who may in fact be upholding some of the poor practice (2020).

## Engaged Communities

Hearteningly, an increase in public awareness, engagement, and activism is helping to drive reform. Google, which was supplying the US military with machine learning software, did not renew its contract with the Pentagon as had been expected after thousands of employees pro- tested at their AI being used to kill via unmanned drones (Changfong-Hagen, 2020; Conger & Cameron, 2018; von Struensee, 2021).

Backlash from the wider public about unethical practices has also led to changes, supported by non-profits who work to protect individual liberty. In the UK in 2020, a court of appeal ruled facial recognition technology use by the Welsh police breached the right to privacy under Article 8 of the European Convention on Human Rights (Wilkinson, 2020). This followed public outcry in 2018 when it was revealed that CCTV cameras in London's King's Cross area were using facial recognition software. In an age when social media can rapidly mobilize support, public engage- ment can be direct and beneficial, using one form of algorithm to campaign against another.

Finally, inclusive practices are taking hold. There are initiatives that call attention to the need for better representation and for marginalized voices to be heard. The AI Now Institute at New York University has produced reports in areas such as gender, race and power, and disability and AI (West et al., 2019b; Whittaker et al., 2019). The Algorithmic Justice League was formed in 2016 by digital activist and computer scientist Joy Buolamwini to raise awareness around AI and racial bias.[3] In her comparatively short career to date, Buolamwini has won a number of prestigious scholarships and her research into algorithmic bias has been

profoundly influential. Since her departure from Google, Timnit Gebru has continued to call out unethical practice and has founded the Distributed Artificial Intelligence Research Institute (DAIR) as a space for "independent, community-rooted AI research."[4]

In terms of marginalized voices, calls to decolonize AI are gaining ground (Mohamed et al., 2020). There is increasing attention to data rights and indigenous data governance, providing a basis for critiquing algorithmic inequality (Carroll et al., 2019). For example, the Canadian non- profit First Nations Information Governance Centre has a data governance strategy that lays out a path to data sovereignty (FNIGC, 2020); the non-profit Open Data Institute (ODI) is running workstreams on data policy and practice focused on structurally under-represented commu- nities;[5] Te Mana Raraunga – the Māori Data Sovereignty Network – advocates for Māori rights and interests in data[6] and was conceived following a workshop in 2015 on the United Nations Declaration on the Rights of Indigenous Peoples.

There is no quick remedy for the discrimination in AI. Despite claims of software solutions that can remove bias, these are limited and cannot be trusted (Gonen & Goldberg, 2019). It is an impossibility to engineer out deep-rooted and systemic historic biases or change the domi- nant culture overnight. For now, we can only seek to make this visible and to mitigate the damage. The power inequalities at a global level – the dominance of corporate control in the US, UK, and Europe – mean that AI systems and their creators are designing and deploying products that inherently benefit them and actively harm the outgroup. When we speak of the exclusion of the global South, Marda explains that, instead of being simply a geographical locale, "The South is rather a metaphor for the human suffering caused by capitalism and colo- nialism on the global level... It is a South that also exists in the geographic North (Europe and North America), in the form of excluded, silenced and marginalised populations, such as undocumented immigrants, the unemployed, ethnic or religious minorities, and victims of sexism, homophobia, racism and Islamophobia" (2019, 12). It is a reminder that intersectional- ity is key when examining the impact of AI. It is a reminder that we must continually ask who is creating this technology, who is benefitting from it, and who is being excluded. Acknowledging the complex power imbalance is the first step and incorporating the voices of the underrepresented is the first fix.

## Notes

1 https://www.image-net.org.
2 https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation. 3 https://www.ajl.org/learn-more.
4 https://www.dair-institute.org.
5 https://theodi.org/article/le-guins-data-identities.
6 https://www.temanararaunga.maori.nz.

## References


Agüera y Arcas, B., Mitchell, M. and Todorov, A. (2017). Physiognomy's New Clothes. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

Anyoha, R. (2017). Blog. *The History of Artificial Intelligence*. Special Edition: Artificial Intelligence. Science in the News, Harvard Graduate School of the Arts and Sciences. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Bartoletti, I. (2018). *An Artificial Revolution: On Power, Politics and AI*. The Indigo Press.

Bell, G. (2021). Touching the future: Stories of systems, serendipity and grace. In *Griffith Review 71: Remaking the Balance*. https://openresearch-repository.anu.edu.au/handle/1885/223254?mode=full

Bilalić, M., Smallbone, K., McLeod, P., & Gobet, F. (2009). Why are (the best) women so good at chess? Participation rates and gender differences in intellectual domains. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1659), 1161-1165. https://doi.org/10.1098/rspb.2008.1576

Bird, S. (2020, December). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3504-3519).

Birhane, A., & Guest, O. (2020). Towards decolonising computational sciences. *arXiv preprint arXiv:2009.14258*. https://arxiv.org/abs/2009.14258

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*, 4349-4357. https://doi.org/10.48550/arXiv.1607.06520

Bones, H., Ford, S., Hendery, R., Richards, K., & Swist, T. (2021). In the Frame: the Language of AI. *Philosophy & Technology*, *34*(1), 23-44. https://doi.org/10.1007/s13347-020-00422-7

Borenstein, J., Grodzinsky, F. S., Howard, A., Miller, K. W., & Wolf, M. J. (2021). AI ethics: A long history and a recent burst of attention. *Computer*, *54*(1), 96–102. https://doi.org/10.1109/MC.2020.3034950

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Burnette, J. (2008). *Gender, Work and Wages in Industrial Revolution Britain* (Cambridge Studies in Economic History - Second Series). Cambridge: Cambridge University Press.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186. https://doi.org/10.1126/science.aal4230

Carlson, N. (2010). At last — the full story of how Facebook was founded. *Business Insider,* 05 March 2010. https://www.businessinsider.com/how-facebook-was-founded-2010-3?r=US&IR=T

Carroll, S. R., Rodriguez-Lonebear, D., & Martinez, A. (2019). Indigenous Data Governance: Strategies from United States Native Nations. https://doi.org/10.5334/dsj-2019-031

Cave, S., Dihal, K., & Dillon, S. (Eds.). (2020). AI Narratives. Oxford University Press.

Changfong-Hagen, K. (2020). " Don't Be Evil": Collective Action and Employee Prosocial Activism. *HRLR Online*, *5*, 188. http://hrlr.law.columbia.edu/hrlr-online/dont-be-evil-collective-action-and-employee-prosocial-activism

Chow, T., & Charles, M. (2019). An negalitarian paradox: On the uneven gendering of computing work around the world. In C. Frieze & J. Quesenberry (Eds.), *Cracking the digital ceiling: Women in computing around the world* (pp. 25–45). Cambridge University Press.

Clark, J. (2016). Artificial Intelligence Has a 'Sea of Dudes' Problem. Bloomberg Professional June 27, 2016

Conger, K., & Cameron, D. (2018). Google is helping the Pentagon build AI for drones. *Gizmodo, March*, *6*.

Costanza-Chock, S. (2020). Introduction: #TravelingWhileTrans, Design Justice, and Escape from the Matrix of Domination. In Design Justice. The MIT Press. https://designjustice.mitpress.mit.edu/

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. *AI & SOCIETY*, 1-12 https://excavating.ai

Daly, A. (2019). Algorithmic oppression with Chinese characteristics: AI against Xinjiang's Uyghurs. APC.

Dechter, R. (1986). Learning While Searching in Constraint-Satisfaction-Problems. *AAAI*.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). IEEE.

Dignum, V. (2019). Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer.

Ensmenger, N. (2015). "Beards, sandals, and other signs of rugged individualism": masculine culture within the computing professions. *Osiris*, *30*(1), 38-65. https://doi.org/10.1086/682955

Ericsson (2021). *Ericsson Mobility Report.* https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/november-2021

Eveleth, R. (2014). How self-tracking apps exclude women. The Atlantic, 15.

Farrow, R. (2019). How an Élite University Research Center Concealed Its Relationship with Jeffrey Epstein. *The New Yorker*, *6*.

First Nations Information Governance Centre (2020). A First Nations Data Governance Strategy. https://fnigc.ca/wp-content/uploads/2020/09/FNIGC_FNDGS_report_EN_FINAL.pdf

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. Science and Engineering Ethics, 26(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Garvie, C., and Frankle, J. (2016). Facial-recognition software might have a racial bias problem. *The Atlantic*, *7*.

Gebru, T. (2020). Race and gender. *The Oxford Handbook of Ethics of AI*, pp. 251-269. https://doi.org/10.1093/oxfordhb/9780190067397.013.16

Gershgorn, D. (2017). The data that transformed AI research—and possibly the world. *Quartz*. 26 July, 2017.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint. https://arxiv.org/abs/1903.03862

Gray, M. L., & Suri, S. (2019). *Ghost Work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt.

Hall, W., Lovegrove, G. (1988). Women and AI. *AI & Soc* **2,** pp. 270–271. https://doi.org/10.1007/BF01908552

Hicks, M. (2017). Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing. The MIT Press.

Hoffmann-Riem, W. (2020). Artificial intelligence as a challenge for law and regulation. In *Regulating artificial intelligence* (pp. 1-29). Springer, Cham.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432. https://doi.org/10.1111/lnc3.12432

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020, July). Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5491-5501.

Kaplan, K. A. (2003). Facemash Creator Survives Ad Board. *The Harvard Crimson*, 19 Nov 2003.

Kelleher, J. D. (2019). *Deep Learning*. MIT press.

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), pp. 1789-1801. http://doi.org/10.1109/TIFS.2012.2214212

Klinger, U., & Svensson, J. (2021). The power of code: women and the making of the digital world. *Information, Communication & Society*, *24*(14), 2075-2090. https://doi.org/10.1080/1369118X.2021.1962947

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097-1105. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Kuczmarski, J. (2018). Reducing gender bias in Google Translate. https://blog.google/products/translate/reducing-gender-bias-google-translate/

Le, Q.V., Ranzato, M., Monga, R., Devin, M., Corrado, G.S., Chen, K., Dean, J., & Ng, A. (2013). Building high-level features using large scale unsupervised learning. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8595-8598.

Lee, D. (2018). Why Big Tech pays poor Kenyans to teach self-driving cars - BBC News. BBC. https://www.bbc.co.uk/news/technology-46055595

Leibold, J. (2020). Surveillance in China's Xinjiang region: Ethnic sorting, coercion, and inducement. *Journal of Contemporary China*, *29*(121), 46-60. https://doi.org/10.1080/10670564.2019.1621529

Leuner, J. (2019). A Replication Study: Machine Learning Models Are Capable of Predicting Sexual Orientation From Facial Images. *arXiv preprint arXiv:1902.10739*. https://arxiv.org/abs/1902.10739

Lighthill, J. (1973). "Artificial Intelligence: A General Survey", Artificial Intelligence: a paper symposium, Science Research Council. http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/ p001.htm

Liveley, G., & Thomas, S. (2020). Homer's Intelligent Machines: AI in Antiquity. In *AI Narratives* (pp. 25-48). Oxford University Press.

Lovelace, A.A. (1843). "Notes" to "A Sketch of the Analytical Engine invented by Charles Babbage, by L.F Menabrea". Notes: vol. 3, pp. 666-731 https://repository.ou.edu/ uuid/6235e086-c11a-56f6-b50d-1b1f5aaa3f5e#page/1/mode/2up

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C.E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133. http://doi.org/10.1007/bf02478259

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*. https://arxiv.org/abs/1801.00631

Mitchell, M. (n.d.) http://www.m-mitchell.com/

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philosophy & Technology. https://doi.org/10.1007/s13347-020-00405-8

Morrison, P., & Morrison, E. (Eds.). (1961). *Charles Babbage on the principles and development of the calculator: and other seminal writings*. Dover Publications.

Murphy, H. (2017). Why Stanford researchers tried to create a 'gaydar'machine. *The New York Times*, *9*.

Newton, C. (2020, December 3). The withering email that got an ethical AI researcher fired at Google. Platformer. https://www.platformer.news/p/the-withering-email-that-got-an-ethical

Noble, S. U. (2013). Google search: Hyper-visibility as a means of rendering black women and girls invisible. *InVisible Culture*, (19).

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E. and Kompatsiaris, I., (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), p.e1356. http://hdl.handle.net/1802/28018

Otis, G. A. And Dillon, N. (2019). Google using dubious tactics to target people with 'darker skin' in facial recognition project. *New York Daily News*, 02 Oct 2019. https://www.nydailynews.com/2019/10/02/google-using-dubious-tactics-to-target-people-with-darker-skin-in-facial-recognition-project-sources/

Peña, P., & Varon, J. (2019). Decolonising AI: A transfeminist approach to data and social justice. In *Global Information Society Watch 2019. Artificial Intelligence: Human rights, social justice and development* (pp. 28–32). APC, ARTICLE 19, and Swedish International Development Cooperation Agency (SIDA)

Plant, S. (1995). The Future Looms: Weaving Women and Cybernetics. *Body & Society*, *1*(3–4), 45–64. https://doi.org/10.1177/1357034X95001003003

Quach, K. (2019). The infamous AI gaydar study was repeated – and, no, code can't tell if you're straight or not just from your face. *The Register*, 5 Mar 2019. https://www.theregister.com/2019/03/05/ai_gaydar/

Ram, A. (2018). UK universities alarmed by poaching of top computer science brains. *The Financial Times*, 9 May 2018. https://www.ft.com/content/895caede-4fad-11e8-a7a9-37318e776bab

Righetti, L., Madhavan, R., & Chatila, R. (2019). Unintended consequences of biased robotic and artificial intelligence systems [ethical, legal, and societal issues]. *IEEE Robotics & Automation Magazine*, *26*(3), 11-13. http://doi.org/10.1109/MRA.2019.2926996

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Rossiter, M.W., 1995. *Before Affirmative Action, 1940-1972*. Johns Hopkins University Press.

Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach, Global Edition 4th. *Foundations*, *19*, 23. https://aima.cs.berkeley.edu/

Schmidt, B. (2015). Rejecting the gender binary: a vector-space operation. *Ben's Bookworm Blog*. http://bookworm.benschmidt.org/posts/2015-10-30-rejecting-the-gender-binary.html

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., & Lockhart, J. W. (2020). Diagnosing Gender Bias in Image Recognition Systems. *Socius*.

Serwaah, P., & Shneor, R. (2021). Women and entrepreneurial finance: a systematic review. *Venture Capital*, 1-29. http://doi.org/10.1080/13691066.2021.2010507

Shaikh, R. (2019). Zuckerberg Revises His "Hot-or-Not" Facebook Origin Story w/ a Noble "Giving People Voice During Iraq War" Fiction. WCCF Tech, 17 Oct 2019.

Shetterly, M. L. (2017). Hidden figures. HarperCollins Nordic.

Simonite, T. (2018a). When It Comes to Gorillas, Google Photos Remains Blind. WIRED, 11 January 2018. https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Simonite, T. (2018b). AI Is the Future—But Where Are the Women? WIRED, 17 August 2018. https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/

Simonite, T. (2021). What really happened when Google ousted Timnit Gebru. *WIRED*, 1-19. https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/

Snow, J. (2018). Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots. *ACLU Northern California blog.* https://www.aclunc.org/blog/amazon-s-face-recognition-falsely-matched-28-members-congress-mugshots

Solmonoff, R. (1956). http://raysolomonoff.com/dartmouth/boxbdart/dart56ray812825who.pdf

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701-1708).

Treviranus, J. (2018). Sidewalk Toronto and Why Smarter is Not Better*. https://medium.datadriveninvestor.com/sidewalk-toronto-and-why-smarter-is-not-better-b233058d01c8

Troiani, I., & Dutson, C. (2021). The neoliberal university as a space to learn/think/work in highereducation. *Architecture and Culture*, 9(1), 5–23. https://doi.org/10.1080/20507828.2021.1898836

Truitt, E. R. Demons and Devices: Artificial and Augmented Intelligence before AI. In *AI Narratives* (pp. 49-71). Oxford University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59,* pp. 433–460. https://doi.org/10.1093/mind/LIX.236.433

US Dept of Labor, Office of Federal Contract Compliance Programs. (2021, February 1). *Google LLC, US Department of Labor settlement resolves alleged pay, hiring discrimination at California, Washington State locations*. News Release Number 21-111-SAN. https://www.dol.gov/newsroom/releases/ofccp/ofccp20210201

von Struensee, S. (2021). The Role of Social Movements, Coalitions, and Workers in Resisting Harmful Artificial Intelligence, and Contributing to the Development of Responsible AI. *Coalitions, and Workers in Resisting Harmful Artificial Intelligence, and Contributing to the Development of Responsible AI (June 16, 2021)*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3880779

West, M., Kraut, R., & Ei Chew, H. (2019a). I'd blush if I could: closing gender divides in digital skills through education.

West, S. M., Whittaker, M., & Crawford, K. (2019b). Discriminating systems: Gender, race and power in AI. *AI Now*.

Whittaker, M., Alper, M., Bennett, C.L., Hendren, S., Kaziunas, L., Mills, M., Morris, M.R., Rankin, J., Rogers, E., Salas, M. and West, S.M. (2019). Disability, bias, and AI. *AI Now Institute*.

Wilkinson, Steve. (2020). "Artificial intelligence, facial recognition technology and data privacy." *Journal of Data Protection & Privacy* 3, no. 2: p. 186-198. https://hstalks.com/article/5408/artificial-intelligence-facial-recognition-technol/

Winfield, A. (2019). Ethical standards in robotics and AI. Nature Electronics, 2(2), 46–48. https://www.nature.com/articles/s41928-019-0213-6

Woolley, B. (2002). *The bride of science: Romance, reason, and Byron's daughter*. New York: McGraw-Hill.

World Economic Forum (2018). *Global Gender Gap Report 2018*. https://www3.weforum.org/docs/WEF_GGGR_2021.pdf