Where are the women?

Mapping the gender job gap in AI

Policy Briefing – Full Report

Erin Young, Judy Wajcman,
Laila Sprejer

The Alan Turing Institute

Public Policy Programme
Women in Data Science
and AI project

Authors

Erin Young is a Postdoctoral Research Fellow in the Public Policy Programme at The Alan Turing Institute. https://www.turing.ac.uk/people/researchers/erin-young

Judy Wajcman is the Principal Investigator of the Women in Data Science and AI project at The Alan Turing Institute, and Anthony Giddens Professor of Sociology at the London School of Economics. https://www.lse.ac.uk/sociology/people/judy-wajcman

Laila Sprejer is a Data Science Research Assistant in the Public Policy Programme at The Alan Turing Institute. https://www.turing.ac.uk/people/researchers/laila-sprejer

Cite as: Young, E., Wajcman, J. and Sprejer, L. (2021). Where are the Women? Mapping the Gender Job Gap in AI. Policy Briefing: Full Report. *The Alan Turing Institute*.

Contents

Introduction

There is a troubling and persistent absence of women employed in the Artificial Intelligence (AI) and data science fields. Over three-quarters of professionals in these fields globally are male (78%); less than a quarter are women (22%) (World Economic Forum, 2018). In the UK, this drops to 20% women. This stark male dominance results in a feedback loop shaping gender bias in AI and machine learning systems.[1] It is also fundamentally an ethical issue of social and economic justice, as well as one of value-in-diversity.[2]

Nearly 4 years ago, the House of Lords Select Committee on Artificial Intelligence (2018) advocated for increasing gender and ethnic diversity amongst AI developers,[3] and last year the European Commission (2020a: 3) noted that it is 'high time to reflect specifically on the interplay between AI and gender equality'. Yet there is still a striking scarcity of quality, disaggregated, intersectional data which is essential to interrogate and tackle inequities in the AI and data science labour force.[4] Indeed, the Royal Society (2019: 51) has noted that 'a significant barrier to improving diversity is the lack of access to data on diversity statistics'. The recent AI Roadmap (UK AI Council, 2021: 4) strongly recommends 'mak[ing] diversity and inclusion a priority [by] forensically tracking levels of diversity to make data-led decisions about where to invest and ensure that underrepresented groups are given equal opportunity'.

As AI becomes ubiquitous in everyday life, closing the gender gap in the AI and data science workforce matters. The fields are particularly fast-moving, so it is important to comprehensively map how these gaps are manifest across different industries, occupations, and skills.

This policy paper is a contribution to this endeavour, presenting a new, curated dataset, analysed through innovative data science methodology, to explore in detail the gendered dynamics of data science and AI careers. This work has added urgency since the drive to close the gender gap in the technology industry risks being derailed by the pandemic. Covid-

[1] See 'The AI Feedback Loop: why diversity matters' below, discussing how the biases of the AI sector are being 'hard-coded' into technologies.

[2] The inclusion of a diverse range of people in the workforce has been shown to boost productivity, profit and innovation (e.g. Herring, 2009; Vasilescu et al., 2015; Tannenbaum et al., 2019).

[3] In 2019 the UK government pledged £13.5 million to fund AI and data science conversion degrees, with 1000 scholarships for people from under-represented groups (Office for Artificial Intelligence, 2019).

[4] Women are a multifaceted and heterogeneous group, with a plurality of experiences, and gender *intersects* with multiple aspects of difference and disadvantage (Crenshaw, 1995; Collins, 1998).

19 is having a disproportionate impact on women across multiple areas, not only exposing but also increasing inequities (Little, 2020; UN Women, 2020; Young, 2020).

As such, this policy briefing from The Alan Turing Institute's Women in Data Science and AI project maps women's participation in data science and AI in the UK and other countries.[5] Our research findings reveal extensive disparities in skills, status, pay, seniority, industry, job, attrition and educational background, which call for effective policy responses if society is to reap the benefits of technological advances.

Our work began with a review of existing statistics and datasets as a baseline. Subsequently, via a partnership with Quotacom, an executive search and consulting firm specialising in data science, advanced analytics and AI, we obtained and analysed a unique dataset which contains career data on individuals working in data fields. This includes links to many of their public LinkedIn profiles. We also present a previously unpublished case study from an innovative review of online global data science platforms.

[5] https://www.turing.ac.uk/research/research-projects/women-data-science-and-ai (Hub: https://www.turing.ac.uk/about-us/equality-diversity-and-inclusion/women-data-science-and-ai)

Key findings

1. **Existing data is sparse:** The existing evidence base about gender diversity in the AI and data science workforce is severely limited. The available data is fragmented, incomplete and inadequate for investigating the career trajectories of women and men in the fields. Where datasets are available, they often rely on commercial data produced through proprietary analyses and methodologies. National labour force statistics lack detailed information about job titles and pay levels within ICT, computing, and technology, which is in particular a major barrier to examining the emerging hierarchy between data science and AI, and other subdomains. These omissions are compounded by a severe lack of intersectional data about the global AI workforce, broken down by age, race, geography, (dis)ability, sexual orientation, socioeconomic status as well as gender. This is particularly concerning since it is those at the intersections of multiple marginalised groups who are at the greatest risk of being discriminated against at work and by resulting AI bias.

2. **Diverging career trajectories:** There is evidence of persistent structural inequality in the data science and AI fields, with the career trajectories of data and AI professionals differentiated by gender. Women are more likely than men to occupy a job associated with less status and pay in the data and AI talent pool, usually within analytics, data preparation and exploration, rather than the more prestigious jobs in engineering and machine learning. This gender skill gap risks stalling innovation and exacerbating gender inequality in economic participation.

3. **Industry differences:** Women in data and AI are under-represented in industries which traditionally entail more technical skills (for example, the Technology/IT sector), and over-represented in industries which entail fewer technical skills (for example, the Healthcare sector). Furthermore, there are fewer women than men in C-suite positions across most industries, and this is even more marked in data and AI jobs in the technology sector.

4. **Job turnover and attrition rates:** Women working in AI and data science in the tech sector have higher turnover (i.e. changing job roles) and attrition rates (i.e. leaving the industry altogether) than men.

5. **Self-reported skills:** Men routinely self-report having more skills than women on LinkedIn. This is consistent across all industries and countries in our sample. This correlates with existing research into women's lower confidence levels in their own technical abilities.

6. **The qualification gap:** Women in data and AI have higher formal educational levels than men across all industries. The achievement gap is even higher for those in more senior ranks (i.e. for C-suite roles), and this ‘over-qualification’ aspect is most marked in the Technology/IT sector. This is particularly striking given that Findings 3 and 5 indicate that women are severely under-represented in the C-suite in the technology industry, and that they self-report having fewer data and AI skills.

7. **Participation in online platforms:** Our research indicates that women comprise only about 17% of participants across the online global data science platforms Data Science Central (‘DS Central’), Kaggle and OpenML. On Stack Overflow, women are a mere 8%. Additionally, we find that only about 20% of UK data and AI researchers on Google Scholar are women. Of the 45 researchers with more than 10,000 citations, only five were women.

Recommendations

1. Reporting standards regarding gender and other workforce characteristics in data science and AI companies urgently need to be developed and implemented. Many of the biggest tech companies provide only headline statistics regarding diversity in their data and AI divisions. Institutions must be more transparent about their workforce and governance diversity. Responsible collection of detailed disaggregated data on women and marginalised groups in these fields must be improved, centrally collated and made available to researchers. This should include data on the proportion, seniority, skills, job tenure, turnover, and remuneration levels of women in the sector, and linked explicitly to issues of bias. The ways in which gender interacts with other sources of inequality such as class, race, ethnicity, religion, disability, age and sexual orientation needs to be a focus of analysis. Governments should apply such reporting requirements to all large tech companies, obliging them to disclose and report on the gender composition of their data science and AI teams.

2. Governments must investigate effective ways to tackle gender data gaps in the AI and data science fields, while maintaining privacy and data protection standards. They should work with national and international organisations to initiate research and advocacy programmes, such as the Inclusive Data Charter (IDC), which promotes more granular data to understand the needs and experiences of the most marginalised in society; the UN Women's Women Count programme, which 'seeks to bring about a radical shift in how gender statistics are used, created and promoted'; and the Data2X project, which aims to improve the 'quality, availability, and use of gender data in order to make a practical difference in the lives of women and girls worldwide'. We recommend working with big technology firms such as LinkedIn that have substantial client databases to begin to build a picture.

3. Countries need to take proactive steps to ensure the inclusion of women and marginalised groups in the design and development of machine learning and AI technologies. For example, the UK government should require companies to scrutinise and disclose the gender composition of their technical, design, management and applied research teams. This must also include mandating responsible gender-sensitive design and implementation of data science research and machine learning. This is an issue of social and economic justice, as well as one of AI ethics and fairness.

4. Given the emerging evidence of biases in AI and discriminatory algorithms, there is an ethical imperative to understand the underlying processes, and to have fair opportunity to

challenge the data, the assumptions, and the metrics employed to mechanise the act of decision-making. We need genuine accountability mechanisms, external to companies and accessible to citizens.

5. Gender inclusive labour market policies, such as paid maternity and parental leave and flexible working hours, must be more effectively implemented and enforced across all industries, and affordable childcare must be provided. These measures are a prerequisite to ensuring that women's disproportionate responsibility for domestic and care work does not inhibit their ability to participate in the digital economy on an equal footing to men. Without them, women will not have equal access to training, re-skilling and job transition pathways, especially in expanding, frontier fields such as data science and AI. This is particularly important given the disproportionate impact of pandemic-related job losses on women.

6. Companies in the tech sector must embed intersectional gender mainstreaming in human resources policy so that women and men are given equal access to well-paid jobs and careers. Actionable incentives, targets and quotas for recruiting, up-skilling, re-training, retaining and promoting women at work should be established, as well as ensuring women's equal participation in 'frontier' technical and leadership roles.

Background

Defining data science and artificial intelligence as a profession

In 2012, Harvard Business Review named data scientist as “the sexiest job of the 21st century.” Yet in actuality, data science is still in its formative period and, as Roca (2019: 3) points out, ‘Artificial Intelligence is not a job title’. Noting the wide array of ways to describe and define data science (and AI) and the associated roles, skills, educational backgrounds, tools and methods,[6] Fayyad and Hamutcu (2020) provide a comprehensive overview of the emergence and current state of data science as a profession. This is important for us to reflect upon, particularly given the speed at which the fields move, in order to delineate the scope of our work at the outset. Whilst we acknowledge that it is still too early to define concretely the fields of data science and AI, the working definitions we use are as such:

Data science: “Using data to achieve specified goals by designing or applying computational methods for inference or prediction” (Fayyad and Hamutcu, 2020)

Artificial Intelligence: “When a machine or system performs tasks that would ordinarily require human (or other biological) brainpower to accomplish” (The Alan Turing Institute, 2021)

Crucially, Berman and Bourne (2015: 1) point out that ‘the emergent field of data science offers the opportunity to narrow the gender gap in STEM… by making diversity a priority early on’. Indeed, we find a very exciting possibility here, as follows. A number of works highlight the role of gender relations in the very definition and gradual configuration of computing more generally as a profession.[7] For example, critiquing the ‘pipeline issue’,[8] feminist historian Hicks (2017: 313) recalls that computer programming was originally the purview of women. However, structural discrimination shifted this, edging women out of the newly prestigious

[6] The UK Government have ‘Data scientist’ guidance - https://www.gov.uk/guidance/data-scientist - and multiple MOOCs, and LinkedIn, similarly suggest ‘career courses’ for becoming a data scientist.

[7] We note that ‘gender’ refers to socio-cultural attitudes, behaviours and identities, and ‘sex’ refers to biological characteristics.

[8] The under-representation of women in the tech sector has traditionally been framed as a ‘pipeline problem’, suggesting that the low numbers of women in tech is due to a low female pool of talent in STEM fields (i.e. because girls are uninterested or lack the skills). However, this perspective neglects technology companies’ failure to attract and retain female talent, shifting the obligation to change onto women (Wajcman, 1991; Hill, Corbett and St. Rose, 2010; Gregg, 2015; Mylavarapu, 2016).

computing jobs.[9] As she explains, 'histories of hidden or devalued computing labour connect powerfully with current trends in information technology and prompt questions about the categories of privilege that silently structure our computing systems today'. What is important to emphasise here is that technical skill is often deployed as a proxy to keep certain groups in positions of power (Abbate, 2012).

As such, a core aim of this report is to re-write the narrative, heightening awareness of the gendered history of computing in order to avoid its replication in AI and data science.

This is particularly important as newly created AI and data science jobs are set to be the well-paid, prestigious and intellectually stimulating jobs of the future. Women and other under-represented groups deserve to have full access to these careers, and to the economic and social capital that comes with them. Further, if the women who do succeed in entering tech are stratified into 'less prestigious' subfields and specialities, rather than obtaining those jobs at the forefront of technical innovation, the gender pay gap will be widened.

Women in AI and data science: what does the existing data tell us?

Women in the tech sector

We begin by presenting a few figures on women in the tech sector as a baseline, before delving into the tech subfields of data science and AI.[10] Firstly, as shown in Figure 1, it is notable that the 15-20% of Computer Science degrees earned by women in the USA (and Western Europe) today is down from nearly 40% in the 1980s (Murray, 2016).[11]

[9] See also Misa (2010), Ensmenger (2012) and Thompson (2019).

[10] Not all countries have the same level of gender (in)equality in their tech workforces. For example, in Malaysia some universities have up to 60% women on computer science programmes, with near parity also reported in some Taiwanese and Thailand institutions (Ong and Leung, 2016).

[11] Indeed, D'Ignazio and Bhargava (2020: 207) point out that 'white men remain overrepresented in data-related fields, even as other STEM (Science, Technology, Engineering and Medicine) fields have managed to narrow their gender gap'.

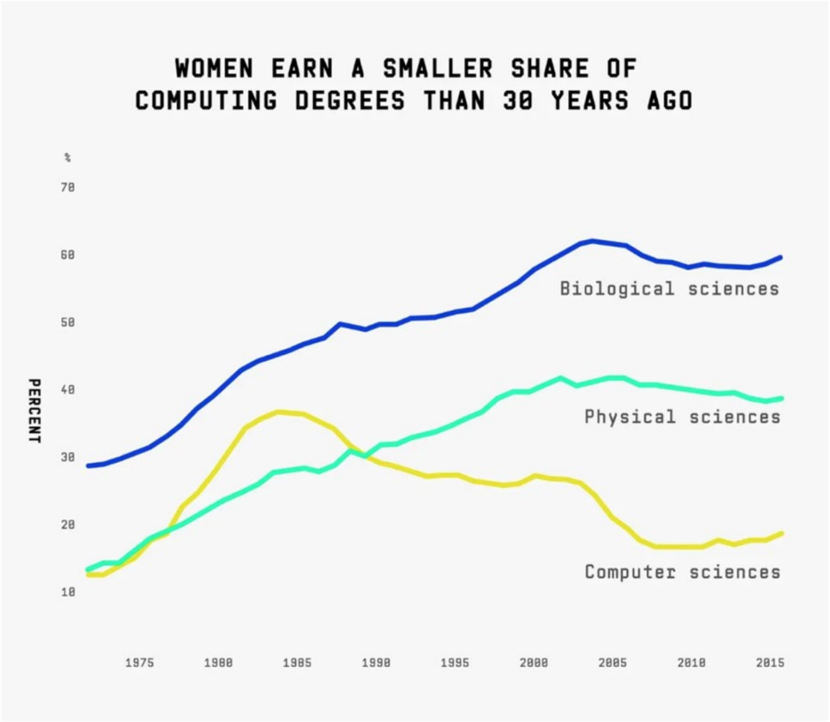


Figure 1: The declining women's share of computer science degrees. Source: National Center for Education Statistics; Chart: WIRED (Simonite, 2018).

According to the 2019 European Commission 'Women in Digital Scoreboard', only 17% of ICT specialists in Europe are women. Similarly, although women make up half the population in the UK, women comprise only ≈17% of its broader technology sector (Inclusive Tech Alliance, 2019). Tech Nation found in 2018 that 19% of UK tech workers were women - notably, this was not reported in the equivalent report in 2020. Additionally, the pay gap in technology fields is estimated to be almost 17% in the UK (Honeypot, 2018).

More recently, the UK tech sector has been found to 'lag behind' in diversity (Goodier, 2020). Indeed, they ranked in 5th place in the Women in Technology Index for the G7 (PwC, 2020: 10). This poor performance on the Index is driven by the UK's worse than average performance on the vast majority of indicators.

Whilst this high-level data exists on the UK tech workforce, it is important to note that, despite acknowledging that 'just one-in-five workers in the technology workforce are female', the 2020 APPG report on Diversity and Inclusion in UK STEM industries does not further segment their data by AI or data science fields. There is thus an urgent need to explore these segments of the tech sector, both in the UK and internationally.[12]

[12] We note that most data on diversity in tech is USA/Europe-centric (and inconsistently collected at that).

Before moving onto existing figures in the AI and data subfields, however, it is also key to highlight the sparsity, but key importance, of intersectional data on the tech sector. Figure 2 illustrates Google's intersectional workforce representation in 2020, but only for the USA. Only 1.6% of Google's US workforce are black women.

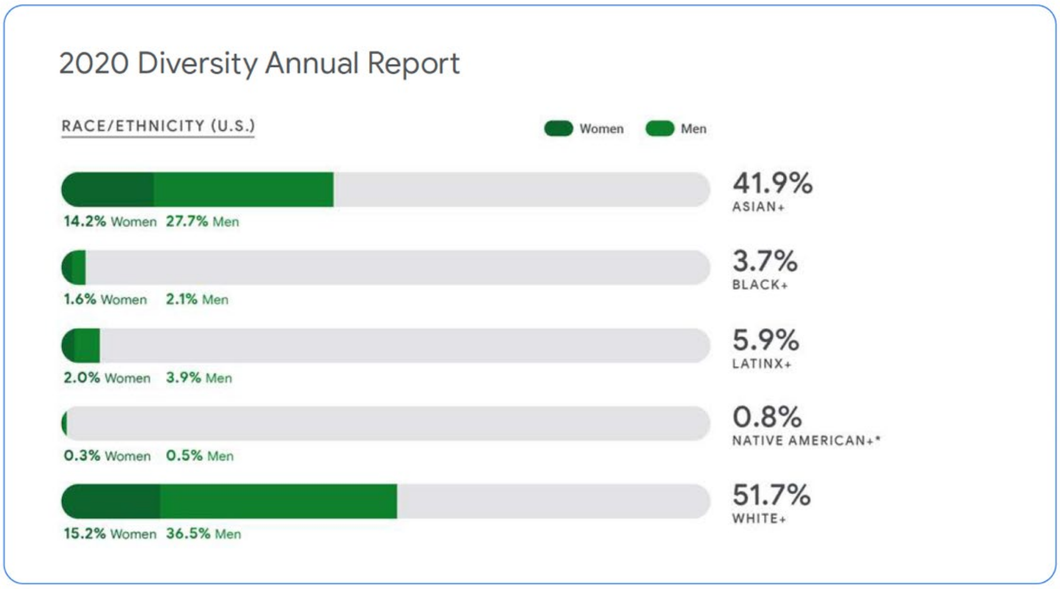
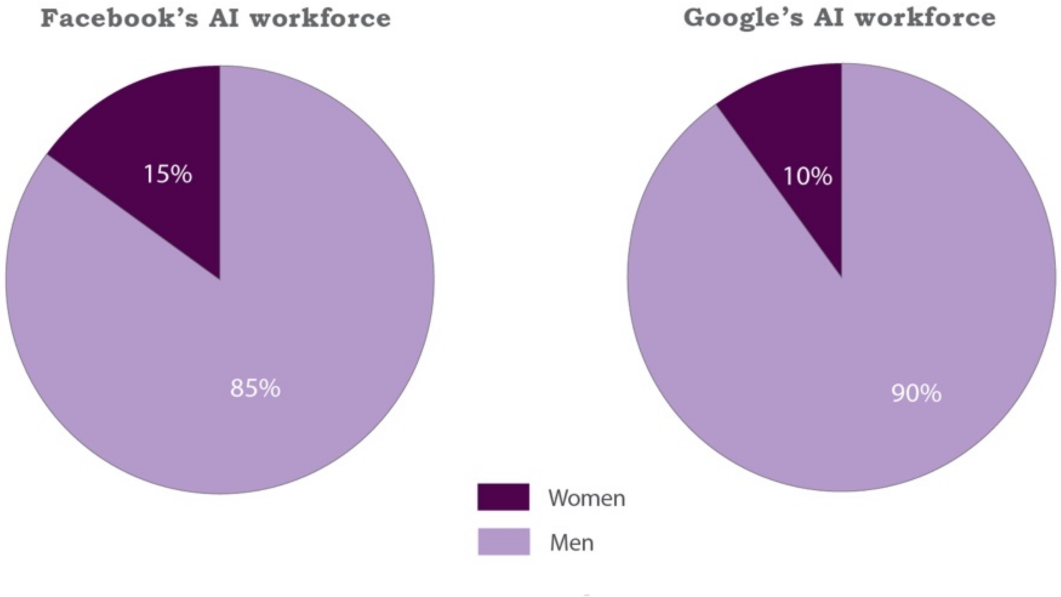


Figure 2: Google's intersectional USA workforce representation. Source: Google Diversity Annual Report 2020.

Disappointingly, Google's Annual Diversity Report 2020 did not show a significant increase from 2019 in the number of women in their workforce, nor in the number of women in leadership roles. Indeed, diversity policies and training (among other initiatives) have only made a marginal difference in growing the share of women in the tech workforce (e.g. Dobbin and Kalev, 2016). As Alegria (2019: 723) explains, 'women, particularly women of colour, remain numerical minorities in tech despite millions of dollars invested in diversity initiatives'.

AI and data science (as subfields of the broader tech sector)

Data specific to the workforce of the tech subfields of data science and AI is much more limited. This is partly because of the lack of clarity in their definitions and the newness of these professions – but, mainly, it is because of an unwillingness of big tech companies to share this data. Indeed, as West, Whittaker and Crawford (2019: 10-12) note, 'the current data on the state of gender diversity in the AI field is dire... [and] the existing data on the state of diversity has real limitations'. They explain that over the past decade, the AI field has shifted from a primarily academic setting to a field increasingly situated in corporate tech environments. 'It is simply harder to gain a clear view of diversity and decision making within the large technology firms that dominate the AI space due to the ways in which they tightly control and shape their hiring data. This is a significant barrier to research... the diversity and inclusion data AI companies release to the public is a partial view, and often contains flaws'. For example, figures 3 and 4 show the extent of Google and Facebook's AI-specific reporting.



Figures 3 and 4: Facebook's and Google's AI workforces, respectively. Sources: Company reported statistics, 2018 (see Simonite, 2018).

There has been some tentatively promising work undertaken by the World Economic Forum in 2018, in collaboration with LinkedIn, exploring gender gaps in AI (see Findings below for discussion). It is important to point out, however, that that unlike the 2018 report, the gender of AI talent is not broken down to the same detail in the more recent World Economic Forum 2020 Global Gender Gap Report. The latter instead only states that women make up 'a relatively lower share of those with disruptive technology skills', comparing the share of men and women in data and AI with other 'professional clusters' (see figure 5).

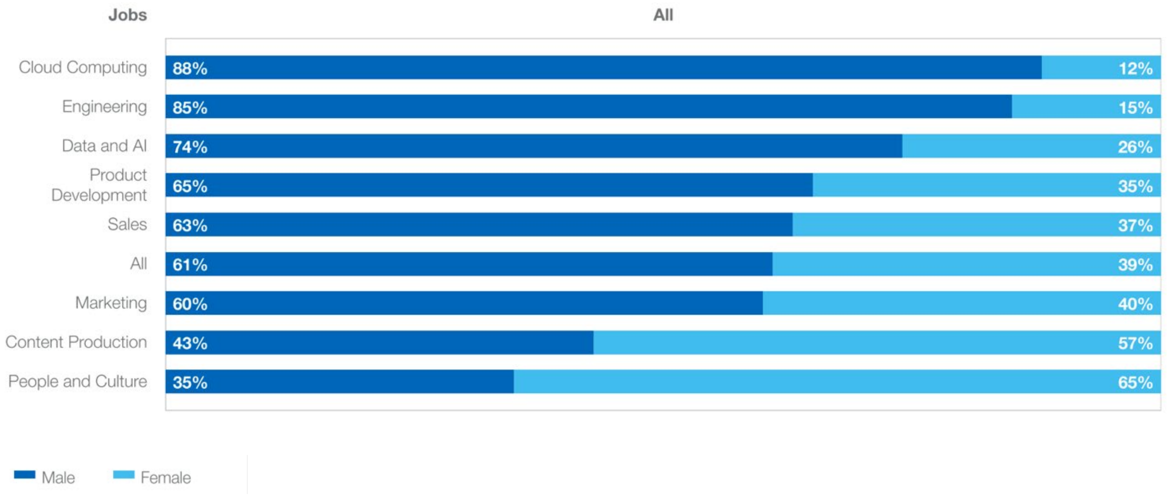


Figure 5: Share of men and women workers across professional clusters. Source: World Economic Forum Global Gender Gap Report 2020.

Comparing the UK statistics with these global figures, we see that there are even fewer women working in the data and AI fields in the UK compared to the global average. Women make up an estimated 26% of workers in data and AI roles globally, which drops to only 22% in the UK. Further, in the UK, the share of women in engineering and cloud computing is a mere 14% and 9% respectively.

Given the scarcity of raw industry data available, researchers have drawn on other sources including online data science platforms (see our case study below), surveys, and academic and conference data (e.g. Freire, Porcaro and Gómez, 2021). These approaches also provide mounting evidence of serious gaps in the gender diversity of the AI research and development workforce. For example, an independent survey of 399 data scientists by the recruiting firm Burtch Works found that 15% were women, although this figure shrank to 10% for those in the most senior roles (Burtch, 2018).

In 2018, WIRED and Element AI reviewed the AI research pages of leading technology companies and found that only 10-15% of machine learning researchers were women (Simonite, 2018). Notably, Google's AI pages listed 641 people working on machine intelligence, but only around 60 were women. Related research found that on average only 12% of authors who had contributed work to the leading three machine learning conferences (NIPS, ICML and ICLR) in 2017 were women (Mantha and Hudson, 2018; Simonite, 2018). This figure drops to 8.19% for the UK specifically (see figure 6).

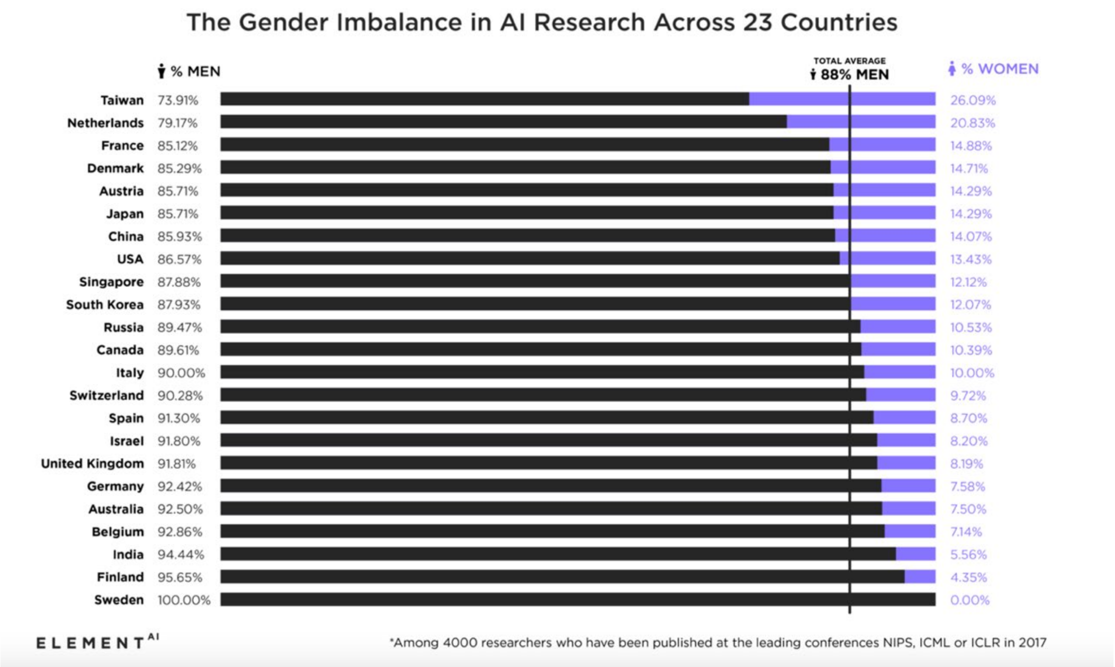


Figure 6: The Gender Imbalance in AI Research across 23 countries. Source: Estimating the Gender Ratio of AI Researchers Around the World (Mantha and Hudson, 2018).

Indeed, there is more information regarding women in AI specifically in research and in the academy, due to the more readily available data. For example, in a large-scale analysis of gender diversity in AI research using publications from arXiv, Stathoulopoulos and Mateos-Garcia (2019) found that only 13.8% of AI paper authors were women. They established that, in relative terms, the proportion of AI papers co-authored by at least one woman has not improved since the 1990s. They also discovered that only 11.3% of Google's researchers who published their AI research on arXiv were women. This proportion was similar for Microsoft (11.95%), and slightly higher, although still low, for IBM (15.66%).

Additionally, the 2019 Artificial Intelligence Index reported that, across all the educational institutions they examined, men constituted a clear majority of AI department faculty, making up 80% of AI professors on average (Perrault et al., 2019). Moreover, diversifying AI faculty along gender lines has not shown significant progress — with women comprising less than 20% of the new faculty hires in 2018. Similarly, the share of female AI PhD recipients has remained virtually constant at 20% since 2010 in the USA.

The statistics and data we have reviewed confirm that the 'newest wings of technology', that is, data science and AI, have dismal representation of women (West, Kraut and Chew, 2019). In other words, the more prestigious and vanguard the field, the fewer the number of women working in it. As the AI and data science fields are rapidly growing as predominant subfields within the tech sector, it seems that so is the pervasive gender gap within them. In order to fully grasp the nature of this problem, we need better data. As the recent AI Index 2021 report stresses:

> *"The lack of publicly available demographic data limits the degree to which statistical analyses can assess the impact of the lack of diversity in the AI workforce on society as well as broader technology development. The diversity issue in AI is well known, and making more data available from both academia and industry is essential to measuring the scale of the problem and addressing it."*

CASE STUDY: DATA SCIENCE AND AI PLATFORM DEMOGRAPHICS

The sparsity of statistics on the demographics of data science and AI professions, particularly in the UK, motivated us to explore other potentially informative sources. As quickly evolving fields in which practitioners need to stay up-to-date with rapidly changing technologies, online communities are an important feature of data science and AI professions. This case study presents a summary of our examination of a selection of online, global data science platforms (Data Science Central, Kaggle, OpenML and Stack Overflow),[13] as well as Google Scholar (UK).[14]

Demographic data were collected from these important platforms. Among the subset of users that had an identifiable binary gender, the estimated proportion of men and women are shown in figure 7 (see Methodological Appendix III for more information). Our research indicates that women are under-represented at a remarkably consistent, and low, 17-18% across the platforms – with Stack Overflow at a much lower 7.9%.

[13] Data Science Central ('DS Central') is a networking site providing an online community for data professionals that comprises blogs, forums and job boards; Kaggle is an informal, gamified framework where users can engage in individual or collaborative data science projects, participate in competitions, and showcase their work; OpenML allows members to share data, code, workflows and wiki contributions; and Stack Overflow is an essential question and answer site for software developers and programmers.

[14] Google Scholar is a database of academic publications on which researchers can create profiles to document and publicise their work. These profiles include details of each author's academic affiliation and citation count. The database is searchable by institution (as indicated by the academic domain name of a user's verified email e.g. '.turing.ac.uk' for a researcher at The Alan Turing Institute) and by field of interest.

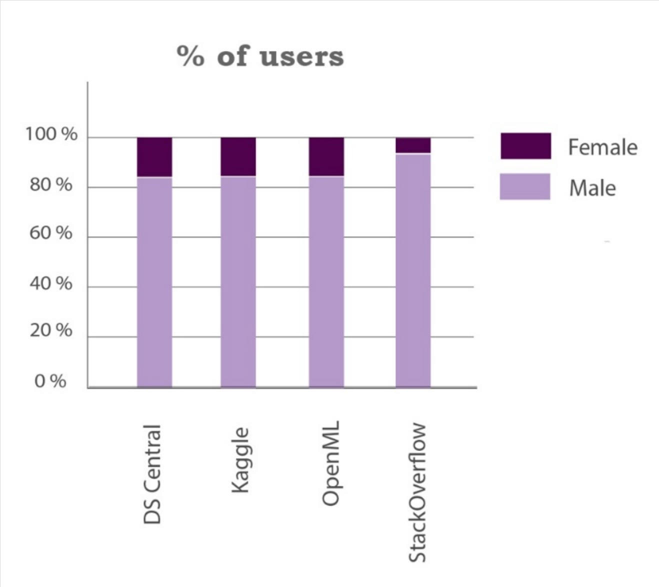


Figure 7: Estimated gender composition of frequently used online data science platforms (May 2019).

Digging further into the Kaggle data, we found that a higher proportion of men have the job titles ‘Software Developer/Engineer’ and ‘Data Scientist’, while a much higher proportion of women have the title ‘Data Analyst’ (see Finding 2 in the main report). Exploring the Data Science Central data, we also found that women are more likely to be employed in the Education and Healthcare sectors, while men are more likely to be employed in Technology and Financial industries (similar to Finding 3). Across all the platforms, women are generally better educated (see Finding 6 in the main report) but worse paid than their male counterparts, and are less likely to have the most prestigious, best-paying job titles. Additionally, the representation of women in data science in the UK is notably poor compared to the USA.

Furthermore, scraping Google Scholar to gather the research profiles of academics across 141 ‘.ac.uk’ domain names, in the fields of AI, machine learning and data science, we find that only 20.2% of such UK researchers with Google Scholar profiles are women. This drops to below 15% among those with the highest citations. Of the 45 researchers with more than 10,000 citations, only five were women.

It is important to note that it is unlikely that any of the platforms considered here mirror exactly the demographics of data scientists and AI professionals as a whole, as these environments will undoubtedly appeal more to some practitioners than others. However, they provide a very interesting lens through which to view participation in the field.

The AI Feedback Loop: why diversity matters

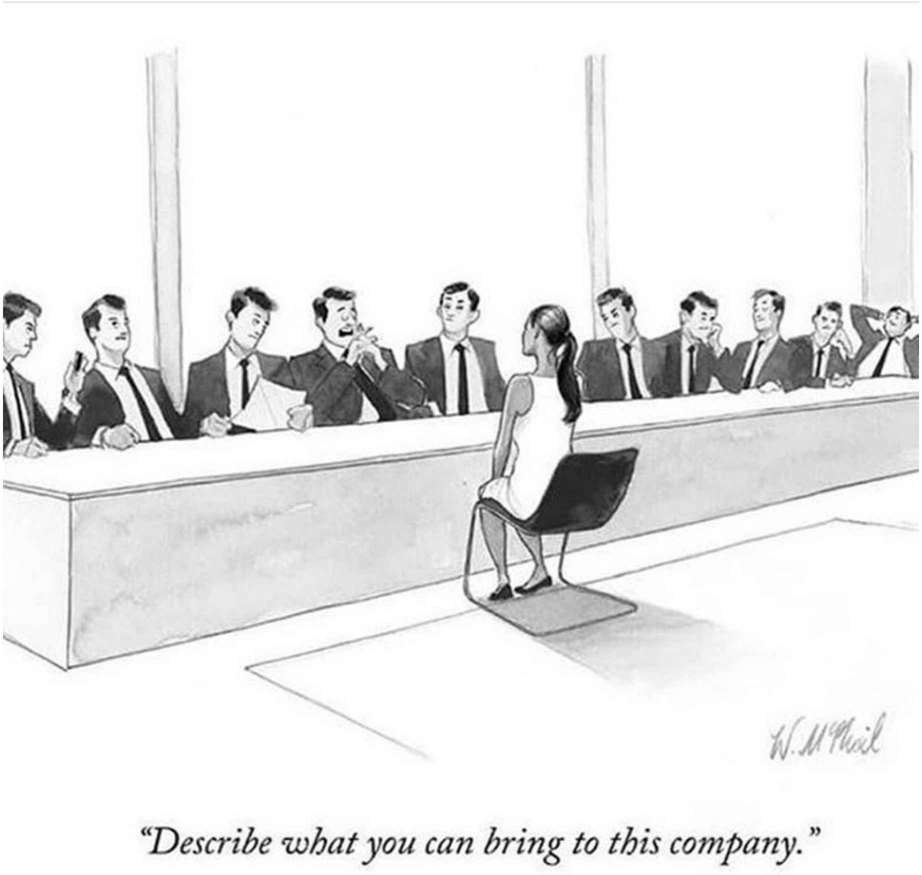


Figure 8: Artist: Will McPhail (The New Yorker).

The stark lack of diversity in the AI and data science fields has wider consequences. Mounting evidence suggests that the under-representation of women in AI results in a feedback loop whereby gender bias gets built into machine learning systems (West, Whittaker and Crawford, 2019; Wajcman, Young and FitzMaurice, 2020).[15] As the European Commission has recognised: 'Technology reflects the values of its developers... It is clear that having more diverse teams working in the development of such technologies might help in identifying biases and prevent them' (Quirós et al., 2018).

Although algorithms and automated decision-making systems are presented and applied as if they are impartial and objective, in fact bias enters, and is amplified through, AI systems at various stages. First, the data used to train algorithms may under-represent certain groups or encode historical bias against marginalised demographics, due to prior decisions on what

[15] See also Leavy (2018), Gebru (2020) and Zacharia et al. (2020).

data to collect, and how it is curated (Criado Perez, 2019; D'Ignazio and Klein, 2020).[16] Second, there are often biases in the modelling or analytical processes due to assumptions or decisions made by developers, either reflecting their own (conscious or unconscious) values and priorities or resulting from a poor understanding of the underlying data. Even the choices behind what AI systems are created can themselves be biased. As O'Neil (2016: 21) succinctly states: 'Models are opinions embedded in mathematics'. If primarily white men are setting AI agendas, it follows that the supposedly 'neutral' technology is bound to be inscribed with masculine preferences (Zou and Schiebinger, 2018).[17]

Several AI products have recently made headlines for their discriminatory outcomes. To name only a few: a hiring algorithm developed by Amazon was found to discriminate against female applicants (Dastin, 2018); a social-media based chatbot had to be shut down after it began spewing racist and sexist hate speech (Kwon and Yun, 2021); the image-generation algorithms OpenAI's iGPT and Google's SimCLR are more likely to autocomplete a cropped photo of a man with a suit, but a woman with a bikini (Steed and Caliskan, 2021; Mahdawi, 2021); and marketing algorithms have disproportionally shown scientific job advertisements to men (Maron, 2018; Lambrecht and Tucker, 2019).[18] The introduction of automated hiring is particularly concerning, as the fewer the number of women employed within the AI sector, the higher the potential for future AI hiring systems to exhibit and reinforce gender bias, and so on.[19]

A number of studies on computer vision have also highlighted encoded biases related to gender, race, ethnicity, sexuality, and other identities (Hendricks et al., 2018; Raji et al., 2020). For instance, facial recognition software successfully identifies the faces of white men but fails to recognise those of dark-skinned women (Buolamwini and Gebru, 2018). Further, research analysing bias in Natural Language Processing (NLP) systems reveal that word embeddings learned automatically from the way words co-occur in large text corpora exhibit

[16] For example, the 'Gendered Innovations 2' report prepared for the European Commission (2020b) found that it is 'possible to introduce bias during the data preparation stage'.

[17] There has been good work by feminist scholars on these issues, such as Eubanks (2018), Noble (2018), Broussard (2018) and Benjamin (2019).

[18] Recently, there has been concern about AI bias in the context of the pandemic (Oertelt-Prigione, 2020). For example, Barsan (2020) found that computer vision models (developed by Google, IBM, and Microsoft) exhibited gender bias when identifying people wearing masks for Covid protection. The models were consistently better at identifying masked men than women and, most worrisome, they were more likely to identify the mask as duct tape, gags or restraints when worn by women.

[19] Similarly, Caliskan, Bryson and Narayanan (2017) show that occupational gender statistics, as we have presented in this report, are 'imprinted' in online text and can be 'mimicked' by machines.

human-like gender biases (Bolukbasi et al., 2016; Gonen and Goldberg, 2019).[20] For example, when translating gender-neutral language related to STEM fields, Google Translate defaulted to male pronouns (Prates, Avelar and Lamb, 2019). Additionally, the common female-gendering of AI voice assistants (such as Siri and Alexa), a deliberate design decision, perpetuate stereotypes of women as obedient, subservient and domestic (Specia, 2019; West, Kraut and Chew, 2019; Yates, 2020; Purtill, 2021).

Finally, it is important to stress that technical bias mitigation (including algorithmic auditing) and fairness metrics for models and datasets are by no means sufficient to resolve bias and discrimination (Foulds et al., 2019; Hutchinson and Mitchell, 2019). Notably, as we elaborate elsewhere (Wajcman, Young and FitzMaurice, 2020), since 'fairness' cannot be mathematically defined, and rather is a political issue, this task often falls to the developers themselves – the very teams in which the diversity crisis lies.

We urgently need more nuanced data and analysis on women in AI in order to better understand these processes and strengthen efforts to avoid hard-coded bias.[21] It is one thing to recall biased technology, but another to ensure that the biased technology is not developed in the first place.[22] As Melinda Gates, Co-chair of the Bill & Melinda Gates Foundation, remarked:

> "If we don't get women and people of colour at the table – real technologists doing the real work – we will bias systems. Trying to reverse that a decade or two from now will be so much more difficult, if not close to impossible" (Hempel, 2017).

[20] See also Garg et al. (2018), Zmigrod et al. (2019) and Strengers et al. (2020).

[21] A curated list of institutions and initiatives tackling bias in AI is available through the Resources section of our Women in Data Science and AI Hub page at https://www.turing.ac.uk/about-us/equality-diversity-and-inclusion/women-data-science-and-ai/resources

[22] West, Kraut and Chew (2019: 88) conclude that 'greater female participation in technology companies does not ensure that the hardware and software these companies produce will be gender-sensitive. Yet this absence of a guarantee should not overshadow evidence showing that more gender-equal tech teams are, on the whole, better positioned to create more gender-equal technology that is also likely to be more profitable and innovative'.

Methodology

We now describe the methodology we employed for our own research, using a novel data science and AI career dataset. In order to gain access to and curate a dataset suitable for investigating (responsibly) gender gaps in these industries, we partnered with Quotacom, an executive search and consulting firm specialising in data science, advanced analytics and AI. From there, we developed a methodology to first identify data profiles, second obtain information on their career trajectories from LinkedIn, and third process the education, work experience and skills into manageable categories. Our purpose was to detect gender gaps across industries as well as general trends around senior women and men working in the data pipeline.

a. Data Collection

Initial seed database

We initially interviewed Quotacom about their data sources and data collection methods in order to understand potential biases in our sample (see Methodological Appendix for details). The Quotacom dataset consists of more than 10,000 'Candidates' (potential recruits) and 90,000 'Contacts' (company contacts), that voluntarily subscribed, either searching for a job or for potential hires. Quotacom scouts across industries, focussing particularly on the data pipeline in EMEA, US and APAC. Data was collected over the last five years, and a GDPR-compliant privacy notice was provided to candidates and contacts before signing up to the database. Each person's job title and LinkedIn profile are provided.

Identifying data and AI profiles

Despite Quotacom's focus on data and AI companies, we found that many 'contacts' in fact did not sit squarely in the data pipeline on which we wanted to focus (those outside of our remit included, for example, non-technical HR administrators, sales executives and account managers). As such, we decided to leverage the database's links to LinkedIn in order to use LinkedIn profiles' job titles as a filter. Since this is a free-text field, after usual pre-processing - i.e., lowercase, stop words removal and stemming – we still had over 40,000 unique job titles to classify. We decided to match these to the International Standard Classification of Occupations (ISCO-08) categorisations from the ILO in order to prevent possible biases from a purely keyword-based approach.[23] First, we used word vectors and similarity scores to find the closest standard title for each profile and its sub-major category, and filtered those within the ILO codes '25 (Information and Communication Technology Professionals)' and '133 (Information and Communication Technology Service Managers)'. To test for biases in our matching we randomly sampled 1,000 profiles and looked for data-related job titles that were misclassified, and added them to the standard ILO job list. We then performed a new matching, this time with an 80% similarity threshold, which left us with 22,373 data profiles. We tested the precision in our detection by randomly sampling 1,000 of the selected profiles and looking to see if the job titles were correctly matched. Out of those, only 92 were wrongly classified (90.8% precision). Similarly, we estimated a 76% recall (i.e. how many data profiles were left out of our sample) by manually validating a random sample of 1,000 profiles from our complete list.

LinkedIn

LinkedIn claims to be *the world's largest professional network with nearly 740 million members in more than 200 countries and territories worldwide*, hosting self-reported information on individual's professional and educational backgrounds and skills. As recognised by Case et al. (2012: 2), 'as a dataset, the LinkedIn database is a valuable information repository'. Similarly, Li et al. (2017) acknowledge that 'given the large-scale digital traces of labour flows available on the web (e.g., LinkedIn), [LinkedIn data] is of considerable interest in understanding the dynamics of employees' career moves'.

Consequently, we decided to scrape LinkedIn to collect the complete educational, professional and skill set information of the individuals on our reduced list of profiles.[24] No personal information, such as phone numbers and email addresses, was collected, and data was fully anonymised in storage.

It is important to note here that the vast majority of LinkedIn information is self-reported and optional. As such, we should keep in mind that some information may be missing, exaggerated, biased towards self-perception, or even subject to different qualification standards (e.g. when stating proficiency in a particular skill). We try to mitigate these by

[23] The ISCO-08 framework provides a means of categorising jobs into different groups according to their tasks and duties. Using their classification system, we matched our 40,000 job titles to their 7,000 titles, and then used their 43 sub-groups to filter IT jobs. Complete details on its structure can be found at: https://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm.
[24] Code for the LinkedIn scraping is available at github.com/sprejerlaila/linkedInScraping/

looking at gender differences in the aggregated data and focusing on the relative gaps rather than the absolute numbers.

b. Data cleaning and characterisation

As stated, one of our major concerns when dealing with LinkedIn data is its level of completeness, especially when each field of information is 'optional'. To ensure a minimum comparability between users, we only considered profiles with some professional experience, and with at least 50 contacts. We also removed outliers according to the years of experience and number of different jobs that they held.[25]

Since all the information collected was filled as free text, there was a significant amount of data cleaning and pre-processing involved before we could start our analysis. A complete description of the variables used, as well as the processing methods, can be found in the Methodological Appendix at the end of this document.

Our final sample consisted of 19,535 profiles, out of which 2,203 (11.3%) are women, belonging mostly to the USA, France, Germany or the UK. Our exploratory analysis showed that, as anticipated by Quotacom, our sample is very senior with an average of almost 20 years of work experience. Further, over 55% of our sample hold a graduate or postgraduate degree (see Table 1a and 1b).

Table 1a and 1b: Characterisation of the sample.

	Female	Male	Graduate degree	Senior jobs
% of total	11.3%	88.7%	55.6%	59.2%
N	2,203	17,332	8,793	10,431

	Years of work experience	Number of different roles	Number of different companies	Number of industries
Mean	19.88	7.32	5.29	3.64
Median	19.83	7	5	3

It is clear that our sample is not representative of the entire global data and AI population. We are aware that our data is not comprehensive, and that it is not intersectional. Rather, we claim that our gender analysis holds for senior profiles who use LinkedIn. Further, in order to account for potential biases in the companies on the Quotacom database, we conduct our analysis at an industry level, and test for prevalence across different countries.

[25] We removed 25 outlier profiles who reported more than 45 total years of experience.

Findings: Gendered careers in data science and AI

1. Existing data is sparse

The existing evidence base about gender diversity in the AI and data science workforce is severely limited, as elaborated above.

2. Diverging career trajectories

There is evidence of persistent structural inequality in the data science and AI fields, with career trajectories (e.g. job segregation and skills specialisations) of data and AI professionals differentiated by gender.

Our research suggests that women are more likely than men to occupy a job associated with less status and pay in the data science and AI talent pool. Figure 9 shows that women have more data preparation and exploration skills, whereas men have more machine learning, big data, general purpose computing (GPC) and computer science skills.[26] The latter are traditionally associated with more prestigious and higher paying careers.

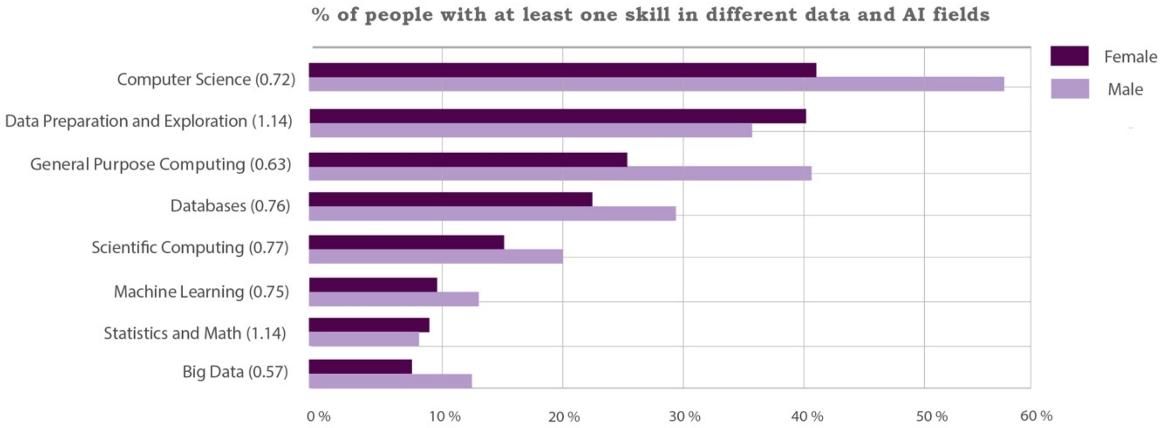


Figure 9: Percentage of people with at least one skill in different data and AI fields. Numbers in brackets represent the gender gap (female/male).

[26] Our 'Statistics and Maths' finding is rather surprising, but we note again that our sample is not statistically representative of the entire population of data and AI professionals.

The most common job field in our dataset for both men and women is Consultancy, with almost no difference by gender.[27] However, consistent with our review of skills, we find that, within the data pipeline, men predominate in Engineering, Architecture and Development jobs, while women do so in Analytics and Research (see Figure 10).

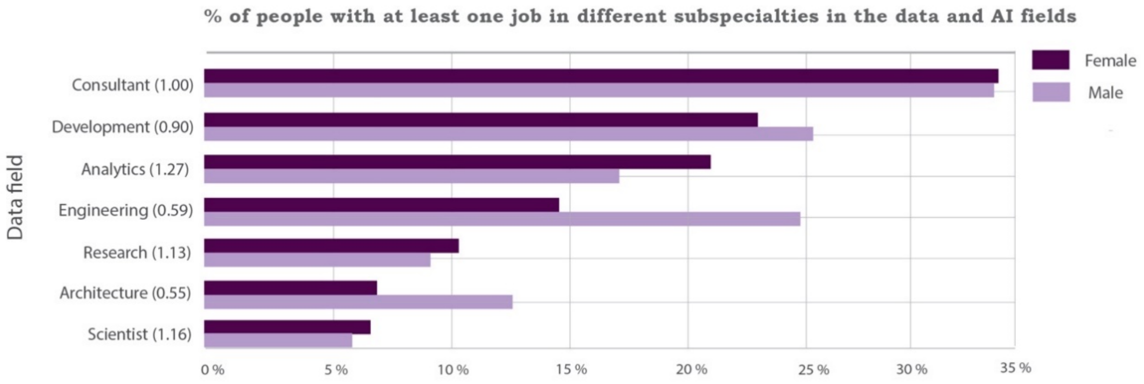


Figure 10: Percentage of people with at least one job in different subspecialities in the data and AI fields. Numbers in brackets represent the gender gap (female/male).

Our findings are consistent with Campero (2021: 62) who found that women are much more prevalent among workers in software quality assurance - crucially, lower-paying and perceived as lower status - than in other software subspecialities. He terms this tendency for women to be segregated into different job subspecialisations than men as 'intra-occupational gender segregation'.[28] Similarly, Guerrier et al. (2009: 506), exploring the gendering of occupational roles within an IT context, note that "women are under-represented in high skilled IT jobs and that a pattern of gender segregation is emerging where women are located in the less technical project management and customer-support roles that are constructed as requiring the sorts of skills that women 'naturally' have". Indeed, as feminist scholars have long evidenced, when women participate in male-dominated occupations, they are often concentrated in the lower-paying and lower-status subfields. 'Throughout history, it has often

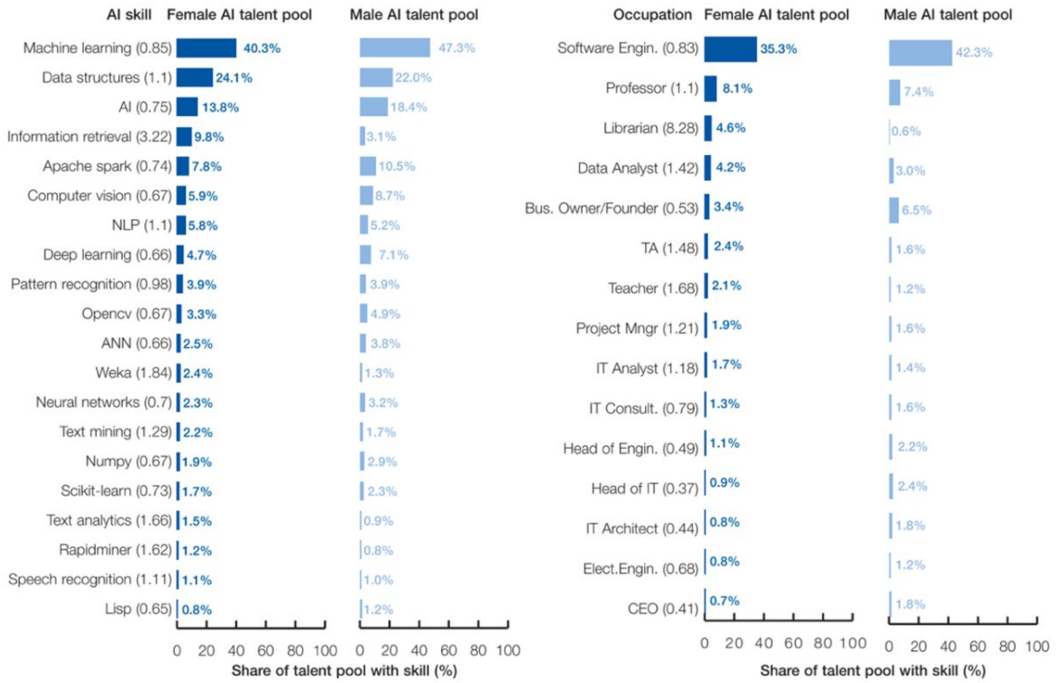
[27] Note: Gender gaps are calculated by dividing % female, by % male. It indicates that, for instance, for every 100 men that report having General Purpose Computing skills, there are only 63 women who do so.

[28] Gender segregation refers to the unequal distribution of men and women in the occupational structure. 'Vertical segregation' describes the clustering of men at the top of occupational hierarchies (higher-paying, higher-status jobs) and of women at the bottom. 'Horizontal segregation' describes the fact that at the same occupational level men and women have different job tasks (see UNESCO, 2020). This is one of the causes of the gender wage gap.

not been the content of the work but the identity of the worker performing it that determined its status' (Hicks, 2017: 16).

As we touched on in our background discussion, as women have begun to enter certain technological subdomains in recent years, such as front-end development, these fields have started to lose prestige and experience salary drops (Posner, 2017; Broad 2019). Meanwhile, men are flocking to the new (prestigious and highly remunerated) data science and AI subspecialities.

Indeed, the Global Gender Gap report (World Economic Forum, 2018) warns about 'emerging gender gaps in Artificial Intelligence-related skills' (see figures 11 and 12). Our results are consistent with their findings that a higher proportion of women than men are data analysts, and higher proportions of men than women are engineers and IT architects. They similarly found that a higher proportion of men have machine learning skills.[29]



Source: LinkedIn.
Note: Gender gaps are indicated in parentheses in the y-axis labels and range from 0 (no women) to 1 (parity). AI = Artificial intelligence, NLP = Natural language processing, ANN = Artificial neural networks, TA = Teaching Assistant, CEO = Chief Executive Officer. ■ = female, ■ = male.

Figures 11 and 12: 'Share of female and male AI talent pool, by AI skill', and 'Share of LinkedIn members with AI skills, by occupation and gender', respectively. Source: World Economic Forum Global Gender Gap report (2018: 31).

[29] Perhaps we are even witnessing the development of a new glass ceiling within the field of Natural Language Processing (NLP), as Schluter's (2018) study suggests.

It is key to note their argument that:

> "AI skills gender gaps may exacerbate gender gaps in economic participation and opportunity in the future as AI encompasses an increasingly in-demand skillset. Second, the AI skills gender gap implies that the use of this general-purpose technology across many fields is being developed without diverse talent, limiting its innovative and inclusive capacity" (World Economic Forum, 2018: viii).

Indeed, there is a hardened talent gap that will require focused intervention. In their recent report proposing elements of a Framework on Gender Equality and AI, UNESCO (2020: 27) point out that 'hiring more women is not enough. The real objective is to make sure that women are hired in core roles such as development and coding'. They recommend the need to substantially increase and bring to positions of parity women coders, developers and decision-makers, with intersectionality in mind. 'This is not a matter of numbers, but also a matter of culture and power, with women actually having the ability to exert influence' (UNESCO, 2020: 23). It is crucial that the AI industry avoid 'participation-washing'; that is, when the mere fact that somebody, here a woman, has participated in a project or endeavour lends it moral and ethical legitimacy (Sloane et al., 2020).[30] Women must have access to the higher status, higher paying roles in the data science and AI fields.

3. Industry differences

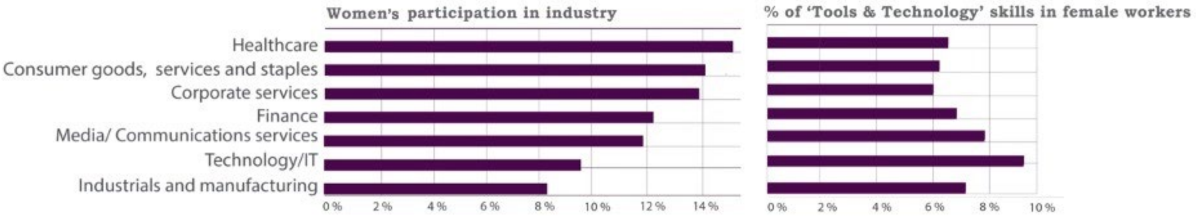
Women in data and AI are under-represented in industries which traditionally entail more technical skills (for example, the Technology/IT sector), and over-represented in industries which entail fewer technical skills (for example, the Healthcare sector). Furthermore, there are fewer women than men in C-suite positions across most industries, and this is even more marked in data and AI jobs in the tech sector.

Our findings suggest that patterns in AI and data science are similar to gender gaps in the overall workforce. Female AI professionals in our sample are more likely to work in 'traditionally feminised' industries which already have a relatively high share of women workers, such as Healthcare. Figure 13 shows that this is also true for the Corporate Services (e.g. Human Resources, marketing and advertising and communications), and Consumer

[30] Mitchell et al. (2020) similarly discuss the difference between heterogeneity in comparison to diversity, with respect to socio-political power disparities.

Goods industries.[31] However, women are under-represented in the Technology/Information Technology (IT) and Industrials and Manufacturing sectors.

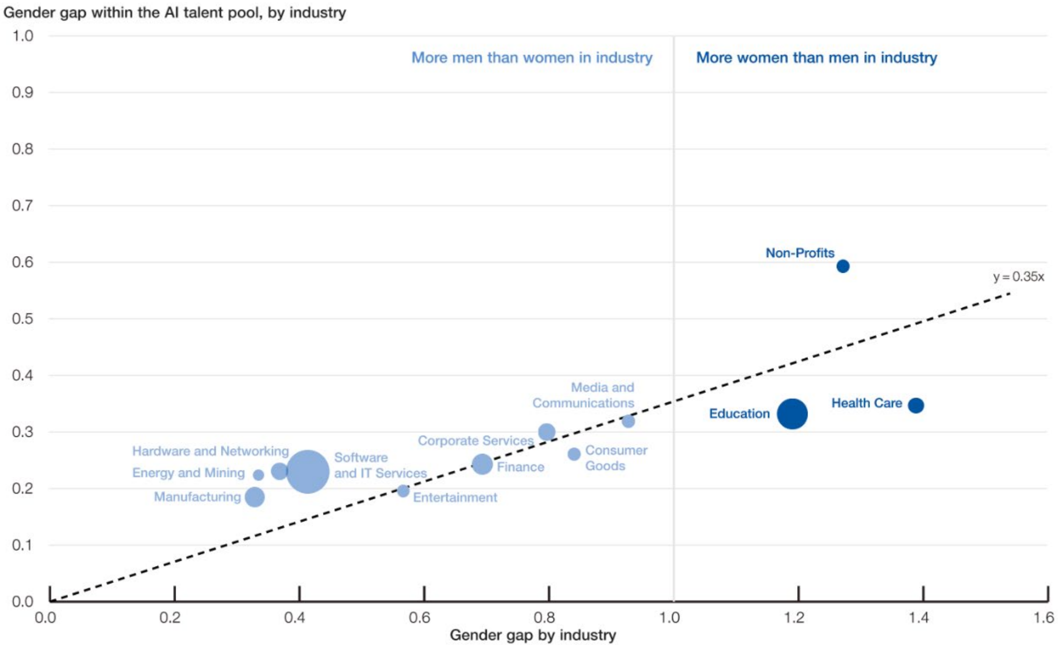
Notably, female participation across different industries is inversely correlated with the percentage of 'Tools and Technologies' skills that they hold (Pearson R of -0.7, p=0.04) (Figure 14). Thus, we found that those industries with lower female participation are also the ones with the higher proportion of 'Tools and Technology' skills in female profiles.



Figures 13 and 14: Women's participation in industry, and % of 'Tools and Technologies' skills held by women workers by industry, respectively. Only industries with a sample of at least 100 female and male profiles each are shown.

Again, our findings are broadly consistent with the World Economic Forum's 2018 Global Gender Gap report. Figure 15, drawn from their report, shows more women than men in the Healthcare industry, and more men than women in the Manufacturing and Software and IT Services sectors.

[31] We were surprised to find a slight over-representation of women in Finance in our sample. However, again, our sample is not representative of the whole population, and we do not intend to provide estimates on overall female participation in industry. Rather, we examine and compare gender gaps within each industry in our sample.



Source: LinkedIn.
Note: Gender gaps range from 0 (no women) to 1 (parity). Size of bubbles represents the size of the AI talent pool. ● = more women than men in industry, ● = more men than women in industry.

Figure 15: 'Gender Gap within the AI talent pool, by industry, across all professionals'. Source: World Economic Forum Global Gender Gap report (2018: 30).

While our data cannot provide evidence for the causation behind this finding, we can confidently speculate as to the reasons why women are under-represented in industries which traditionally entail more technical skills. As we have already noted, stereotypically masculine norms and value systems shape professional practices and career pathways (Muzio and Tomlinson, 2012). These 'masculine defaults', as discussed by Cheryan and Markus (2020), govern technical participation in particular. As Oldenziel (1999) and Miltner (2018) explain, definitions of technological skill and expertise have been historically gendered. They are constructed and framed in such a way that privileges the masculine (as the 'natural' domain of men), rendering the feminine as 'incompatible with technological pursuits' (Wajcman, 2010: 144). Such persistent cultural associations around technology drive women away from, and out of, industries which entail more 'frontier' technical skills such as data science and AI.

It is important to note that we also found a consistent under-representation of women in CXO positions across most industries, regardless of the level of general industry participation (see Figure 16). Even in industries where women are over-represented (for instance, Healthcare), there is still a lower percentage of women in the C-suite.[32]

[32] The exception was 'Media/communication services', which had a higher proportion of women.

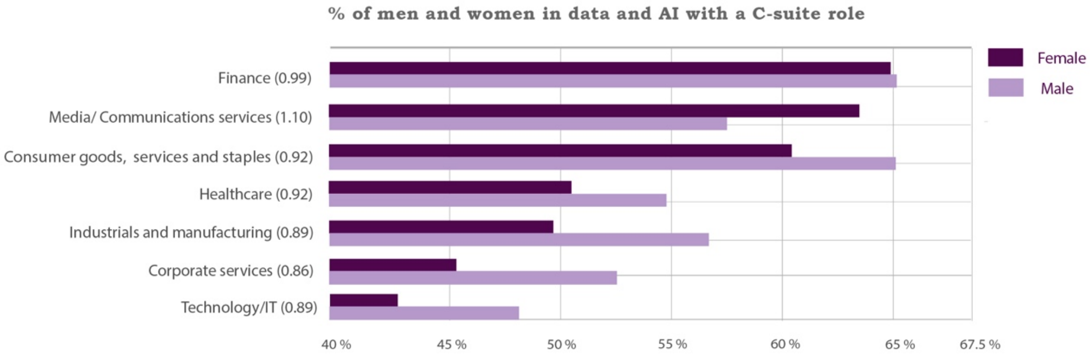
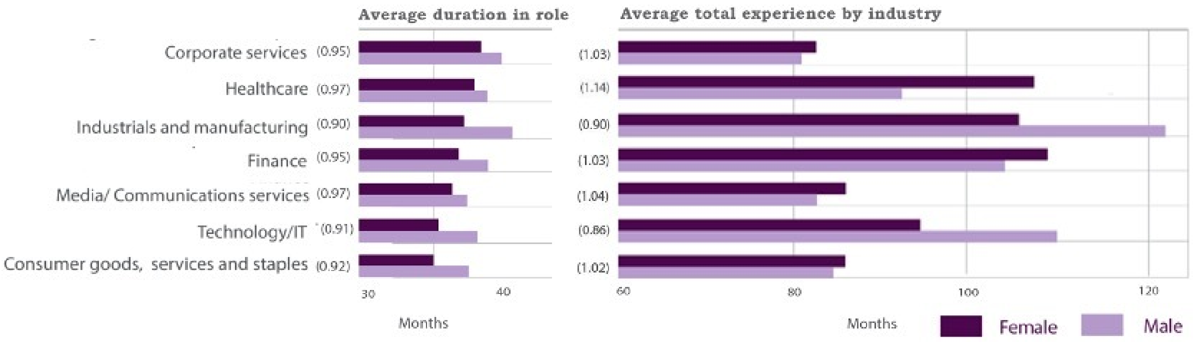


Figure 16: Percentage of men and women in data and AI with a C-suite role, by industry. Numbers in brackets represent the gender gap (female/male).

Best and Modi's (2019) study of women's participation in leadership among top AI companies found that women represent a 'paltry' 18% of C-level leaders among top AI start-ups across much of the globe. They add that of the 95 companies they considered, only two have an equal number of women to men in their C-level positions and none are majority women. Indeed, the World Economic Forum (2018) discovered that, based on LinkedIn data on men and women who hold AI skills, women are less likely to be positioned in senior roles (see Finding 6 below).

4. Job turnover and attrition rates

Women working in AI and data science in the tech sector have higher turnover and attrition rates than men. Like other studies, we have found persistently high turnover (i.e. changing job roles) and attrition rates (i.e. leaving the industry altogether) for women as compared to men working in data science and AI in the technology industry. Our data shows that, on average, women spend less time in each role than men do (see Figure 17). This holds for every industry, with the biggest gap in the Industrials and Manufacturing, and Technology/IT sectors. Furthermore, looking at the total years of experience spent in each industry by gender, we find that on average women spend more time than men in every industry except for Industrials and Manufacturing, and crucially, the Technology/IT sector, where they spend almost a year and a half less (see Figure 18).



Figures 17 and 18: Average duration in role by industry, and average total experience in industry by gender, respectively.

There has been some interesting research on gendered attrition from engineering and technology firms. The US National Centre for Women and Information Technology found that women leave technology jobs at twice the rate of men (Ashcraft, McLain and Eger, 2016). Cardador and Hill (2018) comparably show that women (but not men) taking managerial paths in engineering firms may be at the greatest risk of attrition. In a similar vein, McKinsey found that women made up 37% of entry-level roles in technology, but only 25% reached senior management roles and 15% made executive level (Krivkovich, Lee and Kutcher, 2016).

Exploring the reasons for women's and marginalised groups' high attrition and turnover rates, the Kapor Center argues that unfairness drives turnover, highlighting that 1 in 10 women in technology reported experiencing unwanted sexual attention (Scott, Kapor Klein and Onovakpuri, 2017). Indeed, as other research attests, reasons include 'chilly', unwelcoming environments, workplace discrimination and micro-aggressions,[33] sexual harassment, gendered domestic and family commitments and, as discussed, persistent stereotypes and cultural associations about who 'fits' in technology fields.[34] This is an important aspect which we will explore in our future project work.

5. Self-reported skills

Men routinely self-report having more skills than women on LinkedIn. This is consistent across all industries and countries within our sample.

[33] According to the State of European Tech Survey, 59% of Black/African/Caribbean women have experienced discrimination in some form. An overwhelming 87% of women are challenged by gender discrimination compared to 26% of men (Atomico, 2020).

[34] See, in order, Bobbitt-Zeher (2011), Kolhatkar (2017), Lee (2018), Paul (2019), Maurer and Qureshi (2019), Faulkner (2009), Alfrey and Twine (2016), Margolis and Fisher (2002), Wajcman (2010), and Wynn and Correll (2018).

Our findings suggest that women are more likely to self-report fewer skills than men. Figure 19 shows the distribution of the number of skills reported on LinkedIn grouped by gender. We can see that the whole female distribution is skewed to the left, suggesting that women are less likely to report skills on LinkedIn, compared to men.

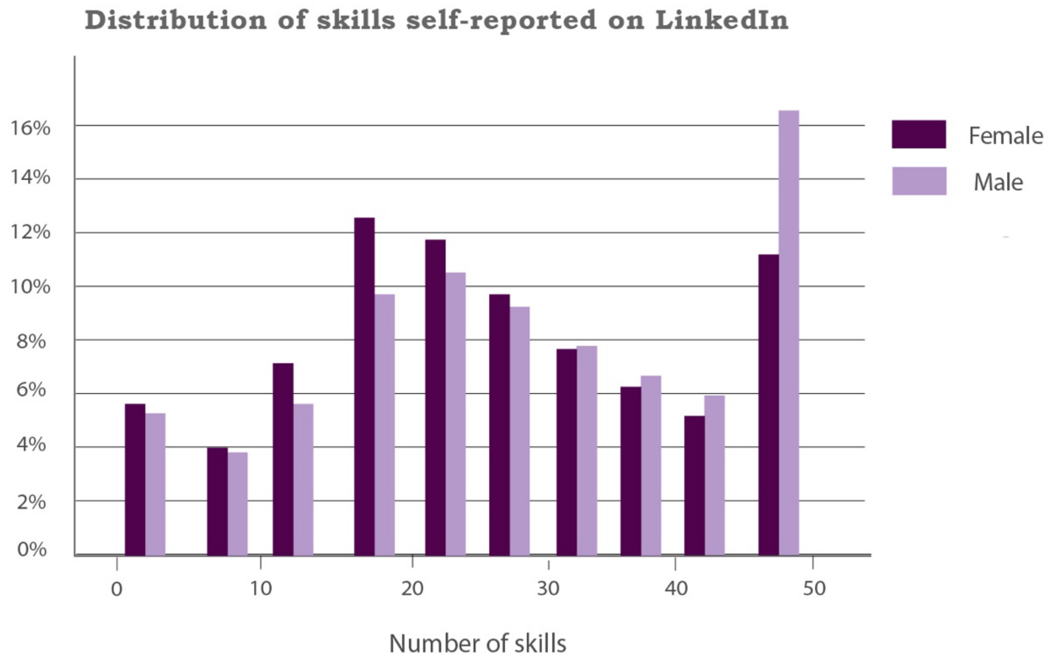


Figure 19: Distribution of the number of skills self-reported on LinkedIn, by gender.

Our findings echo those of Stanford's Human-Centered Artificial Intelligence Institute (HAI), who also tentatively explored the gendering of AI skills using LinkedIn data in their 2019 AI Index report (Perrault et al., 2019). They found that, across all countries, men tended to report AI skills across more occupations than women.[35] Further, referencing the 2018 Global Gender Gap report, Duke (2018) notes that there are '...no signs that this gap is closing: over the past four years, men and women have been adding AI skills to their [LinkedIn] profiles at a similar rate. This means that while women aren't falling further behind, they also aren't catching up'.

Indeed, other studies have also found that women are more modest than men in expressing their accomplishments, and are less self-promoting (Lerchenmueller, Sorenson and Jena, 2019). They also indicate that women are generally less confident in their own abilities, particularly during self-assessment (Correll, 2001; Cech et al., 2011). As touched upon earlier, persistent cultural associations around femininity as 'incompatible' with advanced technological pursuits (alongside 'brogrammer' stereotypes and 'hustling', for example)

[35] It is interesting to note they also found that the UK performs poorly with regards to diversity in comparison to a number of other countries (see Background and Case Study above).

affect women's confidence in their technical skills, shaping perceptions of their aptitude and proficiencies (Jacobs, 2018).

Altenburger et al. (2017: 463) take this point further to speculate as to how these gender differences in self-assessment and self-presentation might affect online professional opportunities, for example on LinkedIn.[36] Women's less favourable assessments of their abilities, fit and belonging in male-dominated data science and AI occupations may well be influential in determining women's aspirations in these fields.[37]

6. The qualification gap

Women in data and AI have higher formal educational levels than men across all industries. The achievement gap is even higher for those in more senior ranks (i.e. for C-suite roles), and this 'over-qualification' aspect is most marked in the Technology/IT sector.

We find that 59% of women in our sample hold a graduate (or postgraduate) degree, compared to 55% of men. This trend also holds when the sample is broken down by industry. Further, when we compared the formal educational levels of our whole sample with a subsample of the most senior profiles (see Figure 20), we found that the educational gap is even higher for those at C-Suite level.

In fact, the gap is roughly double in every industry; by which we mean that, for instance, in all Technology/IT roles, there is an achievement gap of 6%, but for CXO roles, this shoots up to 13%. In the case of the Technology/IT industry, the leap is mostly explained by an increase in the percentage of graduate women in the C-suite. This strongly suggests that women are educating themselves in order to get promoted, while men may not be doing so. The finding is in line with existing thought that women have to work harder and need more qualifications than men in order to progress into senior ranks in the workplace (Scott, 2021).

[36] This could be an interesting further consideration in relation to how the 'pipeline problem' is framed.

[37] See also Leslie et al. (2015) and Wynn and Correll (2017).

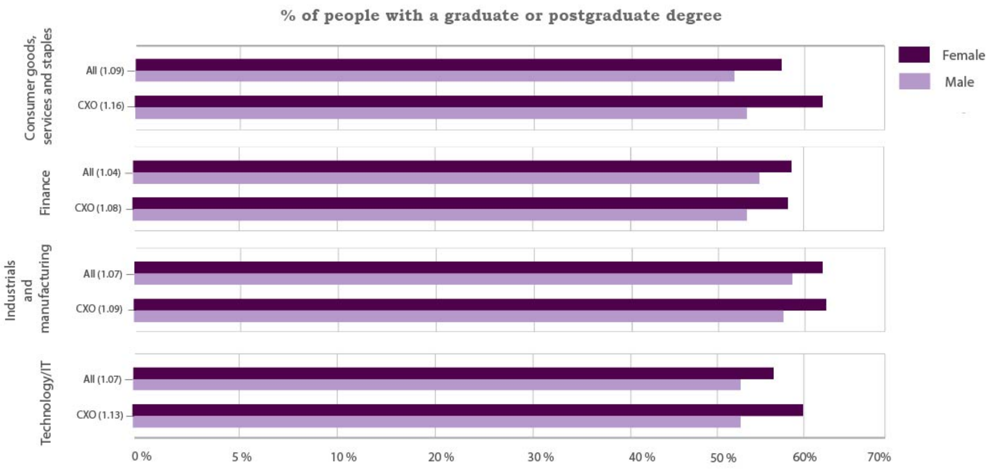


Figure 20: Percentage of men and women with a graduate or postgraduate degree across the whole sample, and across the subsample of C-Suite individuals. Numbers in brackets represent the gender gap (female/male).

This finding is particularly striking given that findings 3 and 5 indicate that women are severely under-represented in the C-suite in the technology industry, and that they self-report having fewer data and AI skills.

Conclusions

Our research, based on a unique dataset of AI professionals, indicates that data science and AI careers in the UK and globally are heavily gendered. There is persistent structural inequality in these fields associated with extensive disparities in skills, status, pay, seniority, industry, attrition rates, educational background, and even self-confidence levels. This gender job gap needs rectifying so that women can fully participate in the AI workforce, including in powerful leadership roles in the design and development of AI.

Our findings are consistent with existing work on the AI gender gap. They require urgent attention given the disproportionate impact of the Covid-19 pandemic on women which risks widening the gender gap in the tech industry (Little, 2020). As Leavy (2018: 16) says: 'advancing women's careers in the area of Artificial Intelligence is not only a right in itself; it is essential to prevent advances in gender equality supported by decades of feminist thought being undone'.[38]

This is not only about issues of economic opportunity and social justice, but also crucially about AI innovation, fairness and ethics. As evidence mounts of gender, race and other social biases embedded in algorithms, there is the risk that AI systems will amplify existing inequities. We cannot even begin to remedy this, let alone take advantage of the huge potential of AI, without first having a data and AI workforce who are representative of the people those systems are meant to serve.

Whilst it is clear that there is a worrying lack of women in the data science and AI fields, there is a scarcity of detailed, intersectional, publicly available demographic information about the data and AI workforce. This is primarily due to the unwillingness of large technology firms to disclose their own diversity data. The lack of transparency has serious implications for Government policymaking around technological advancement and equity, and for labour market policies.[39] It is crucial that we develop a better understanding of the dynamics of the

[38] Similarly, Kumpula-Natri and Regner (2020) argue that 'improving female involvement, and advocating equality and non-discrimination as fundamental principles for developing artificial intelligence, are among the most important feminist objectives of the 2020s'.

[39] 'To ensure that the professions of the future can target gender parity within the coming decade, reskilling and up skilling efforts for women interested in expanding their skills range should be focused on those already in the labour market or looking to re-enter the labour market after a period

problem. This policy report, in both summary and full form, provides a first step in building a robust evidence base to comprehend the dearth of women working in such fields, and its relationship with biased AI. In our future work, the Alan Turing Institute's Women in Data Science and AI project will build upon this research in order to explore the factors driving the AI gender gap.

of inactivity. In tandem, a rigorous diversity and inclusion agenda within organizations can direct hiring practices to fully utilise existing talent pools and ensure that inclusive working environments retain and develop the women already employed in frontier professions' (World Economic Forum, 2020b: 42).

Methodological Appendix

I. Quotacom data collection

Understanding the sources and methods from which our initial data seed list was created is crucial to ensure robustness in our findings. By interviewing Quotacom, we learned that their database profiles were identified and collected in a number of ways. The company creates talent lists for candidates through the use of X-Ray, Lusha, Owler, Skrapp, LinkedIn/Xing or similar, personal networks, referrals, recommendations, websites, industry forums, blogs, competitions, speaker lists, conference attendee lists and industry press. They then approach the candidates via Internet-sourced contact details. Alternatively, candidates can approach Quotacom via responses to advertisements, although they are not typically added to the database unless they have relevant skills within digital transformation, data, data science or AI. Contact profiles - that is, individuals based at Quotacom's partner companies - are sourced in a slightly different way. Companies are initially added as target prospects, and Quotacom then perform various outreach marketing campaigns to stakeholders within those companies, usually via email, phone and LinkedIn. Quotacom typically use LinkedIn, business directories, CrunchBase and Google to develop the initial prospect companies lists. These 'prospects' span from small to large companies, and there are no specific criteria apart from the fact that they operate or have specialist business units in Digital Transformation or Data. Once prospect lists are compiled in Excel, they are loaded onto the central database via CSV.

II. List of variables and data processing

This section describes our complete list of variables, along with their sources and the processing steps taken for this analysis.

Table 2: Complete list of variables and their sources.

Variable	Source	Description
LinkedIn profile	Quotacom	LinkedIn URL
Gender	Genderize API	Inferred gender (binary)
Job history	LinkedIn	Includes: self-declared job title, company, industry, and years.
------- Seniority	Own authors	Inferred from job title based on keywords
------- Role (e.g. consultant, engineer, analyst)	Own authors	Inferred from job title based on keywords
------- Industry	LinkedIn	Industry associated with each job company

------- Start and end date	LinkedIn	Start and end date of employment
Education history	LinkedIn	Includes: self-declared degree, discipline, institution and years.
------- Max degree	Own authors	Maximum degree achieved. Classified into undergrad, post-grad and none.
Skills (LinkedIn)	LinkedIn	List of self-declared skills and their LinkedIn categories.
------- Data skills	Own authors	Subset of data skills and their category based on Fayyad and Hamutcu (2020)
Location	LinkedIn	Inferred location based on their last job.

Gender

In order to infer each profile's gender, first names were passed to an API that returns a gender with a probability score (Genderize API).[40] This method is imperfect as it assumes that gender is binary, and can be inferred from name alone, which is not the case. However, if we are interested in how people are treated because of their perceived gender, this is a reasonable approximation to make, and one that has been widely employed in the literature when studying gendered behaviour online (e.g. Karimi et al., 2016; Terrell et al., 2017; Stathoulopoulos and Mateos-Garcia, 2019). After obtaining the scores for all available names, we manually reviewed scores of less than 0.8 and removed the ones we could not classify (less than 1%). This left us with 11% women in our data.

Location

We used the last available job location for each profile to determine their country of residence, with help from *pycountry* in cases where the name of the city or the country code was mentioned instead of the country name. We found that 50% of our sample corresponded to the US (22.7%), France (10.7%), Germany (10.1%) and the UK (9%), with no significant differences in gender gaps between them in the years of experience, roles duration or number of skills.[41]

[40] https://genderize.io/. The API is designed to predict the gender probability of a person given their name, and is based on more than 100M datapoints collected over 242 countries.

[41] The other 50% is divided between 14 countries that make up 1-5% of the sample each, and 50+ more countries at under 1% each.

Work experience, seniority and fields

For each profile, we scraped their available job history including job title, start and end dates, company name, industry, and location. Using each role duration, we estimated the total years of experience within the same company, industry, and along their whole careers. Further, we used last available work experience to infer our sample's seniority by classifying job titles into five different categories (see Table 3).

Table 3: Keywords used for seniority classification.

Seniority	Keywords
Junior	Junior, Assistant, Intern, Trainee, Associate
Mid	Lead, Manager, Supervisor, Project director
Senior	Senior, Executive, Director, Head, Principal
CXO	Chief X Officer, CXO
Board	VP, President, Chairman, Board, Founder, Partner, Owner

As anticipated by our interview with Quotacom, we found that our sample is very senior, with over 50% having CXO roles, as well as a trajectory of 20 years of experience across 7 different roles (see Table 4).

Table 4: Work experience statistics after removing outliers with more than 3.5 standard deviations over the mean years of experience.

	Total years of experience	Number of different roles	Number of different companies	Number of industries
Mean	19.88	7.32	5.29	3.64
Sd	7.22	2.53	2.43	1.71
Min	1.00	1	1	1
25%	14.75	6	4	2
50%	19.83	7	5	3
75%	24.42	9	7	5
Max	45.33	17	14	13

Finally, we looked at the different job fields by classifying all job titles into Consultancy, Engineering, Development, Analytics, Architecture, Science and Research. We should note that this categorisation was only made for Junior, Mid-, and some Senior roles, given that generally this does not make sense for CXO and Board roles.

Industry

When available, we used the industry from each company's LinkedIn page associated with our profiles' jobs. We then grouped 147 unique LinkedIn industry codes into 13 major categories (see Table 5) and looked at the gender distribution of roles in each one (Figure 21).

Table 5: List of industries and their categorisation.

Industry (from LinkedIn)	Industry group	Industry (from LinkedIn)	Industry group	Industry (from LinkedIn)	Industry group
[All universities]	Academia	Glass, Ceramics & Concrete	Energy & materials	Airlines/Aviation	Industrials and manufacturing
Animation	Arts	Mining & Metals	Energy & materials	Automotive	Industrials and manufacturing
Arts & Crafts	Arts	Oil & Energy	Energy & materials	Aviation & Aerospace	Industrials and manufacturing
Fine Art	Arts	Packaging & Containers	Energy & materials	Civil Engineering	Industrials and manufacturing
Graphic Design	Arts	Paper & Forest Products	Energy & materials	Computer Hardware	Industrials and manufacturing
Music	Arts	Plastics	Energy & materials	Construction	Industrials and manufacturing
Performing Arts	Arts	Renewables & Environment	Energy & materials	Electrical & Electronic Manufacturing	Industrials and manufacturing
Photography	Arts	Semiconductors	Energy & materials	Import & Export	Industrials and manufacturing
Apparel & Fashion	Consumer goods	Accounting	Finance	Industrial Automation	Industrials and manufacturing
Business Supplies & Equipment	Consumer goods	Banking	Finance	Machinery	Industrials and manufacturing
Consumer Electronics	Consumer goods	Capital Markets	Finance	Maritime	Industrials and manufacturing
Cosmetics	Consumer goods	Financial Services	Finance	Mechanical Or Industrial Engineering	Industrials and manufacturing
Dairy	Consumer goods	Insurance	Finance	Medical Device	Industrials and manufacturing
Farming	Consumer goods	Investment Banking	Finance	Package/Freight Delivery	Industrials and manufacturing
Fishery	Consumer goods	Venture Capital & Private Equity	Finance	Railroad Manufacture	Industrials and manufacturing
Food & Beverages	Consumer goods	Investment Management	Finance	Shipbuilding	Industrials and manufacturing
Food Production	Consumer goods	Alternative Dispute Resolution	Government, NGOs & Legislation	Transportation/Trucking/Railroad	Industrials and manufacturing
Furniture	Consumer goods	Civic & Social Organization	Government, NGOs & Legislation	Warehousing	Industrials and manufacturing

Gambling & Casinos	Consumer goods	Defense & Space	Government, NGOs & Legislation	Building Materials	Industrials and manufacturing
Leisure, Travel & Tourism	Consumer goods	Environmental Services	Government, NGOs & Legislation	Broadcast Media	Media/communications services
Luxury Goods & Jewelry	Consumer goods	Executive Office	Government, NGOs & Legislation	Consumer Services	Media/communications services
Ranching	Consumer goods	Fundraising	Government, NGOs & Legislation	Entertainment	Media/communications services
Recreational Facilities & Services	Consumer goods	Government Administration	Government, NGOs & Legislation	Information Services	Media/communications services
Restaurants	Consumer goods	Government Relations	Government, NGOs & Legislation	Media Production	Media/communications services
Retail	Consumer goods	Individual & Family Services	Government, NGOs & Legislation	Motion Pictures & Film	Media/communications services
Sporting Goods	Consumer goods	International Trade and Development	Government, NGOs & Legislation	Newspapers	Media/communications services
Sports	Consumer goods	Judiciary	Government, NGOs & Legislation	Online Media	Media/communications services
Supermarkets	Consumer goods	Law Enforcement	Government, NGOs & Legislation	Printing	Media/communications services
Textiles	Consumer goods	Law Practice	Government, NGOs & Legislation	Publishing	Media/communications services
Tobacco	Consumer goods	Legal Services	Government, NGOs & Legislation	Telecommunications	Media/communications services
Utilities	Consumer goods	Legislative Office	Government, NGOs & Legislation	Writing & Editing	Media/communications services
Wholesale	Consumer goods	Military	Government, NGOs & Legislation	Architecture & Planning	Other
Wine & Spirits	Consumer goods	Non-profit Organization Management	Government, NGOs & Legislation	Commercial Real Estate	Other
Consumer Goods	Consumer goods	Philanthropy	Government, NGOs & Legislation	Design	Other

Events Services	Corporate Services	Public Policy	Government, NGOs & Legislation	Libraries	Other
Facilities Services	Corporate Services	Public Safety	Government, NGOs & Legislation	Program Development	Other
Human Resources	Corporate Services	Security & Investigations	Government, NGOs & Legislation	Real Estate	Other
Logistics & Supply Chain	Corporate Services	Think Tanks	Government, NGOs & Legislation	Religious Institutions	Other
Management Consulting	Corporate Services	Translation & Localization	Government, NGOs & Legislation	Biotechnology	Technology/IT
Market Research	Corporate Services	International Affairs	Government, NGOs & Legislation	Computer & Network Security	Technology/IT
Marketing & Advertising	Corporate Services	Museums & Institutions	Government, NGOs & Legislation	Computer Networking	Technology/IT
Outsourcing/Offshoring	Corporate Services	Political Organization	Government, NGOs & Legislation	Computer Software	Technology/IT
Public Relations & Communications	Corporate Services	Alternative Medicine	Healthcare	Information Technology & Services	Technology/IT
Staffing & Recruiting	Corporate Services	Health, Wellness & Fitness	Healthcare	Internet	Technology/IT
E-learning	Education	Medical Practice	Healthcare	Mobile Games	Technology/IT
Education Management	Education	Mental Health Care	Healthcare	Nanotechnology	Technology/IT
Higher Education	Education	Pharmaceuticals	Healthcare	Computer Games	Technology/IT
Professional Training & Coaching	Education	Veterinary	Healthcare	Wireless	Technology/IT
Research	Education	Hospital & Health Care	Healthcare		
Chemicals	Energy & materials	Hospitality	Healthcare		

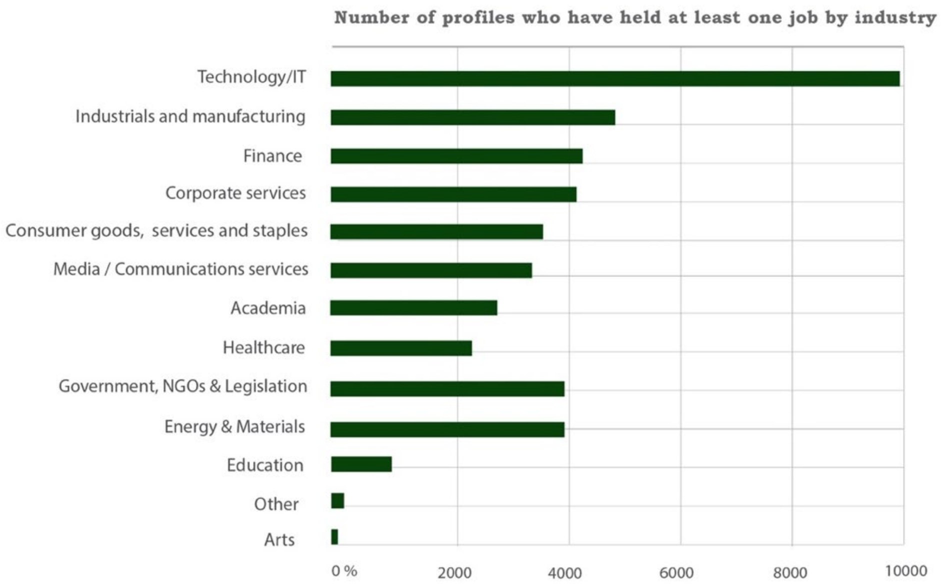


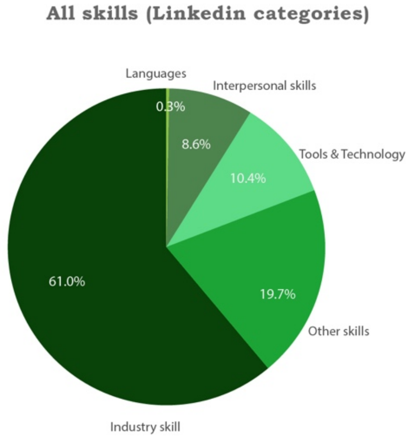
Figure 21: Number of profiles who have held at least one job by industry.

Figure 21 shows the number of profiles that have held at least one job in each industry. Unsurprisingly, Technology/IT is the most common, with over 50% of the individuals having worked in a tech company.

Skills

LinkedIn allows users to add up to 50 skills, and automatically classifies them into one of five categories: Industry Knowledge, Tools & Technologies, Interpersonal Skills, Languages, and Other Skills. We found that 7% of our sample had no skills on their LinkedIn, with little difference by gender (6.9% for men and 7.4% for women). For the rest, Figure 22 shows that the prevalence of the types of skills is very uneven, with industry skills encompassing over 60% of the sample.

In order to specifically detect Data Science and AI skills, we used the framework proposed by Fayyad and Hamutcu (2020) by which we re-classified all skills by adding eight new data categories. In our overall sample, data skills represent 15% of the total, and are distributed as shown in Figure 23.



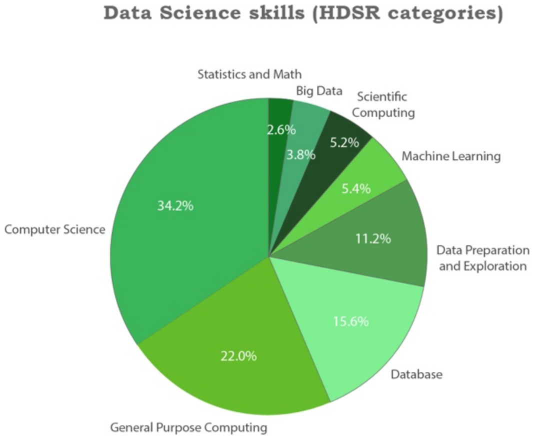


Figure 22: Distribution of skills, as classified by LinkedIn across the whole sample.

Figure 23: Distribution of data science skills, as classified by Fayyad and Hamutcu (2020).

III. Case study

DS Central had a total of 127,678 registered users. The profiles of a third of these (42,204 users) were scraped to obtain details about their gender, location, job title, and interests. Of the 91% of users who listed a binary gender on their profile, **18.1%** identified as female.

In 2017, **Kaggle** conducted a user survey, which received 16,716 responses, asking multiple choice questions about users' demographics, their experiences in data science, and their use of the platform. In total, 16.6% of survey respondents identified as female, 81.4% identified as male, and 2% identified as other (non-binary, genderqueer, gender non-conforming, or a different identity). This corresponds to **17.0%** of users with a binary gender identifying as female.

Note: For the Kaggle survey data, 25% of respondents lived in the US, 16% in India, 3% in Russia, and 3% in the UK. 45% were aged 22-30. Further, the more recent 2020 Kaggle report 'State of Machine Learning and Data Science' reports 16.4% women on their platform, with only 0.3% of people identifying as non-binary.

OpenML had 7126 registered accounts. For this report, these account names were scraped and gender inferred from them. Of the 6153 users for whom a binary gender could be determined, **17.0%** were women.

Note: Inferred gender from first names using the Genderize API (described earlier). A total of 1638 unique profiles were returned by Google Scholar for machine learning, AI, and data science researchers in the UK. Of these profiles, a gender was identified for 88.9% (1456 profiles). Among this subset of researchers, 20.2% of profiles belong to women.

Each year, **Stack Overflow** conducts a user survey; the 2019 survey had nearly 90,000 respondents, of whom 6460 identified themselves as having a speciality in data science or machine learning. Within this subset, of the 6142 respondents that listed a binary gender, **7.9%** identified as female.

Acknowledgements

We would like to thank **Quotacom** for providing the seed dataset for our research. Quotacom is an international executive search and consulting firm with expertise in the digital transformation and data domains.[42] With offices in Europe and the USA, Quotacom is recognised as the leading recruitment specialist within decision science. Quotacom prides itself on diversity and inclusion and has a strong focus on women in technology, with over 60% of the team being female, including a number of members of the senior leadership team. The company hold long-term, strategic partnerships with their clients, ranging from VC backed start-ups, consulting firms and Fortune 500 enterprises, focussing on digital transformation through frontier technologies across Data, Advanced Analytics, AI, Robotics, Machine Learning, Open Source, IOT, Cloud and Blockchain.

We would also like to thank **Dr Anna FitzMaurice** (Senior Data Scientist, BBC) for her work on the Case Study (Data Science and AI platform demographics) in this report.

[42] https://www.quotacom.com/

References

Abbate, J. (2012). Recoding Gender: Women's Changing Participation in Computing. MIT Press.

Alegria, S. (2019). Escalator or Step Stool? Gendered Labor and Token Processes in Tech Work. Gender & Society, 33 (5), 722–745.

Alfrey, L. and Twine, F. W. (2017). Gender-fluid geek girls: Negotiating inequality in the tech industry. Gender & Society, 31 (1), 28 – 50.

All-Party Parliamentary Group (APPG) on Diversity and Inclusion in STEM (2020). The State of the Sector: Diversity and representation in STEM industries in the UK. Data Analysis Brief. Inquiry into the STEM Workforce. British Science Association.

Altenburger, K. M., De, R., Frazier, K., Avteniev, N. and Hamilton, J. (2017). Are there gender differences in professional self-promotion? An empirical case study of LinkedIn profiles among recent MBA graduates. Proceedings of the 11th International Conference on Web and Social Media (ICWSM 2017), October, 460–463.

Ashcraft, C., McLain, B. and Eger, E. (2016). Women in tech: the facts. National Center for Women & Technology (NCWIT).

Atomico (2020). 'Diversity & Inclusion'. In: The State of European Tech 2020. Retrieved from: https://2020.stateofeuropeantech.com/chapter/diversity-inclusion/article/diversity-inclusion/

Barsan, I. (2020). Research Reveals Inherent AI Gender Bias: Quantifying the accuracy of vision/facial recognition on identifying PPE masks. Wunderman Thompson. Retrieved from: https://www.wundermanthompson.com/insight/ai-and-gender-bias

Benjamin, R. (2019). Race After Technology: Abolitionist Tools for the New Jim Code. Cambridge, MA: Polity Press.

Berman, F. D. and Bourne, P. E. (2015). Let's make gender diversity in data science a priority right from the start. PLoS Biology, 13 (7), 1–5.

Bernal, N. (2019). UK AI gender diversity is in 'crisis' as number of female scientists drops. The Telegraph (July 17). Retrieved from: https://www.telegraph.co.uk/technology/2019/07/17/uk-ai-gender-diversity-crisis-number-female-scientists-drops/

Best, M. L. and Modi, D. (2019). ‘Case Study 4: Where are the Women? Gender disparities in AI research and development’, page 16. In: A. Sey and N. Hafkin (2019). Taking stock: data and evidence on gender equality in digital access, skills, and leadership. Preliminary findings of a review by the EQUALS Research Group. EQUALS Global Partnership, United Nations University.

Bobbitt-Zeher, D. (2011). Gender discrimination at work: Connecting gender stereotypes, institutional policies, and gender composition of workplace. Gender and Society, 25 (6), 764–786.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. and Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embedding. 30th Conference on Neural Information Processing Systems, (NIPS 2016).

Broad, E. (2019). Computer says no. Inside Story (April 29). Retrieved from: https://insidestory.org.au/computer-says-no/

Broussard, M. (2018). Artificial Unintelligence: How Computers Misunderstand the World. Cambridge, MA: MIT Press.

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency.

Burtch, L. (2018). The Burtch Works Study Salaries of Data Scientists. Retrieved from: https://www.burtchworks.com/wp-content/uploads/2018/05/Burtch-Works-Study_DS-2018.pdf

Caliskan, A., Bryson, J. and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356, 183-186.

Campero, S. (2021). Hiring and Intra-occupational Gender Segregation in Software Engineering. American Sociological Review.

Cardador, M. T. and Hill, P. L. (2018). Career Paths in Engineering Firms: Gendered Patterns and Implications. Journal of Career Assessment, 26 (1), 95–110.

Case, T., Gardiner, A., Rutner, P. and Dyer, J. (2012). A LinkedIn analysis of career paths of information. Journal of the Southern Association for Information Systems, 1 (1), 1–13.

Cech, E., Rubineau, B., Silbey, S. and Seron, C. (2011). Professional role confidence and gendered persistence in engineering. American Sociological Review, 76 (5), 641–666.

Cheryan, S. and Markus, H. R. (2020). Masculine defaults: Identifying and mitigating hidden cultural biases. Psychological Review, 127 (6), 1022–1052.

Collins, P. H. (1998). It's All In the Family: Intersections of Gender, Race, and Nation. Hypatia, 13(3), 62–82.

Correll, S. J. (2001). Gender and the Career Choice Process: The Role of Biased Self-Assessments. American Journal of Sociology, 106: 1691–730.

Crenshaw, K. W. (1995). Mapping the margins: Intersectionality, identity politics, and violence against women of color. In: K. Crenshaw, N. Gotanda, G. Peller and K. Thomas. (Eds.) Critical race theory: The key writings that formed the movement New York: New Press.

Criado Perez, C. (2019). Invisible Women: Exposing Data Bias in a World Designed for Men. Chatto & Windus.

Dastin, J. (2018). "Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women." Reuters (10 October). Retrieved from: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Data2X. (n.d.). "Important Data about Women and Girls Is Incomplete or Missing." Retrieved from: https://data2x.org

Data Science Central (n.d.). 'Data Science Central: A community for Big Data practitioners'. Retrieved from: https://www.datasciencecentral.com/

Davenport, T. H. and Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review. Retrieved from: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century de Vassimon Manela, D., Errington, D., Fisher, T., van Breugel, B., and Minervini, P. (2021). Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models. ArXiv.

D'Ignazio, C. and Bhargava, R. (2020). "Data Visualization literacy: A feminist starting point." In: M. Engebretsen and H. Kennedy (Eds.) Data Visualization in Society. Amsterdam University Press.

D'Ignazio, C. and Klein, L. F. (2020). Data Feminism. Cambridge, MA: MIT Press.

Dobbin, F. and Kalev, A. (2016). "Why Diversity Programs Fail". Human Resource Management, Harvard Business Review, July-August. Retrieved from: https://hbr.org/2016/07/why-diversity-programs-fail

Duke, S. (2018). Growing But Not Gaining: Are AI skills holding women back in the workplace? LinkedIn Economic Graph (December 18). Retrieved from: https://economicgraph.linkedin.com/blog/growing-but-not-gaining-are-ai-skills-holding-women-back-in-the-workplace

Element AI. (2019). Global AI talent report. Element AI. Retrieved from: https://jfgagne.ai/talent-2019/

Ensmenger, N. L. (2012). The Computer Boys Take Over: Computers, Programmers, and the Politics of Technical Expertise. MIT Press.

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York, NY: St Martin's Press.

European Commission (2019). 'Women in Digital Scoreboard'. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/women-digital-scoreboard-2019-country-reports#:~:text=The%20%22Women%20in%20Digital%22%20scoreboard,employment%20based%20on%2013%20indicators

European Commission (2020a). Opinion on Artificial Intelligence – opportunities and challenges for gender equality. Advisory Committee on Equal Opportunities for Women and Men. (18 March). Retrieved from: https://ec.europa.eu/info/publications/list-previous-opinions-advisory-committee-equal-opportunities-women-and-men-2014-2016_en

European Commission (2020b). Gendered Innovations 2: How Inclusive Analysis

Contributes to Research and Innovation. Policy Review. H2020 Expert Group to update and expand ‘Gendered Innovations/ Innovation through Gender’. Luxembourg: Publications Office of the European Union.

Faulkner, W. (2009). Doing gender in engineering workplace cultures. I. Observations from the field. Engineering Studies, 1 (1), 3–18.

Fayyad, U. and Hamutcu, H. (2020). Toward Foundations for Data Science and Analytics: A Knowledge Framework for Professional Standards. Harvard Data Science Review, 2 (2). Retrieved from: https://hdsr.mitpress.mit.edu/pub/6wx0qmkl/release/3

Foulds, J., Islam, R., Keya, K. N. and Pan, S. (2019). ‘An Intersectional Definition of Fairness.’ arXiv:1807.08362.

Freire, A., Porcaro, L. and Gómez, E. (2021). Measuring Diversity of Artificial Intelligence Conferences. Proceedings of Machine Learning Research 1,10, 2021 AAAI Workshop on Diversity in Artificial Intelligence (AIDBEI 2021).

Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. PNAS, 115 (16).

Gebru, T. (2020). Race and Gender. In: M. Dubber, F. Pasquale and S. Das (Eds.) (2020). The Oxford Handbook of Ethics of AI. Oxford, UK.

Global Partnership for Sustainable Development Data (n.d.). “Inclusive Data Charter.” Retrieved from: https://www.data4sdgs.org/inclusivedatacharter

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. NAACL HLT: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 609–614.

Goodier, M. (2020). UK tech sector “lagging behind” in diversity as study reveals only one in five company officers are female. NS Tech (20 July). Retrieved from: https://tech.newstatesman.com/business/uk-tech-sector-diversity-study

Google (2020). Google Diversity Annual Report 2020. Google.

Google Scholar (n.d.). ‘Google Scholar’. Retrieved from: https://scholar.google.com/

Gregg, M. (2015). The Deficiencies of Tech's 'Pipeline' Metaphor. The Atlantic (December 3). Retrieved from: https://www.theatlantic.com/business/archive/2015/12/pipeline-stem/418647/

Guerrier, Y., Evans, C., Glover, J. and Wilson, C. (2009). “Technical, but not very.”: Constructing gendered identities in IT-related employment. Work, Employment and Society, 23 (3), 494–511.

Hao, K. (2021). An AI saw a cropped photo of AOC. It autocompleted her wearing a bikini. MIT Technology Review. Retrieved from: https://www.technologyreview.com/2021/01/29/1017065/ai-image-generation-is-racist-sexist/

Hempel, J. (2017). Melinda Gates and Fei-Fei Li Want to Liberate AI from “Guys With Hoodies”. WIRED (May 4). Retrieved from: https://www.wired.com/2017/05/melinda-gates-and-fei-fei-li-want-to-liberate-ai-from-guys-with-hoodies/

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T. and Rohrbach, A. (2018). Women Also Snowboard: Overcoming Bias in Captioning Models. ICML Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2018), Stockholm, Sweden.

Hern, A. (2019). Apple made Siri deflect questions on feminism, leaked papers reveal. The Guardian (September 6). Retrieved from: https://www.theguardian.com/technology/2019/sep/06/apple-rewrote-siri-to-deflect-questions-about-feminism

Herring, C. (2009). Does Diversity Pay?: Race, Gender, and the Business Case for Diversity. American Sociological Review, 74 (2): 208-224.

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

Hill, C., Corbett, C. and St. Rose, A. (2010). Why So Few? Women in Science, Technology, Engineering and Mathematics. American Association of University Women (AAUW).

Honeypot (2018). Women in Tech Index, 1–6. Retrieved from: https://www.honeypot.io/women-in-tech-2018/

House of Lords (2018). AI in the UK: ready, willing and able? Report of Session 2017–19. Select Committee on Artificial Intelligence. HL Paper 100. Authority of the House of Lords.

Hutchinson, B. and Mitchell, M. (2019). 50 Years of Test (Un)fairness: Lessons for Machine Learning. FAT* '19, January 29–31, Atlanta, GA, USA.

Inclusive Tech Alliance (2019). "Women in Tech Briefing: A new call to arms." Retrieved from: https://www.inclusivetechalliance.co.uk/

International Labour Organization (ILO) (2016). International Standard Classification of Occupations (ISCO), ISCO-08 Structure, index correspondence with ISCO-88. Retrieved from: https://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm

International Labour Organization (ILO) (2020). ILO Monitor: COVID-19 and the world of work. Fifth edition. Updated estimates and analysis.

Jacobs, J. (2018). Macho 'brogrammer' culture still nudging women out of tech. Financial Times (December 10). Retrieved from: https://www.ft.com/content/5dd12c50-dd41-11e8-b173-ebef6ab1374a

Kaggle (n.d.). 'kaggle'. Retrieved from: https://www.kaggle.com/

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M. and Strohmaier, M. (2016). Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods, WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada, 8–9.

Kolhatkar, S. (2017). The Tech Industry's Gender Discrimination Problem. The New Yorker. Retrieved from: https://www.newyorker.com/magazine/2017/11/20/the-tech-industrys-gender-discrimination-problem

Krivkovich, A., Lee, L. and Kutcher, E. (2016). Breaking down the gender challenge. McKinsey. Retrieved from: https://www.mckinsey.com/business-functions/organization/our-insights/breaking-down-the-gender-challenge

Kumpula-Natri, M. and Regner, E. (2020). Letter: In search for feminist AI, representation is queen. The Financial Times (December 23). Retrieved from: https://www.ft.com/content/9c4d6e7c-2ce8-43fe-b52a-a0343fac4ba2

Kwon, J. and Yun, H. (2021). AI Chatbot Shut Down After Learning to Talk Like a Racist Asshole. Vice World News. Retrieved from: https://www.vice.com/amp/en/article/akd4g5/ai-chatbot-shut-down-after-learning-to-talk-like-a-racist-asshole?__twitter_impression=true

Lambrecht, A. and Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Management Science, 65, 2966-2981.

Leavy, S. (2018). Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. In: GE '18: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, 14-16. New York, NY: Association for Computing Machinery.

Lee, D. (2018) Google staff walk out over women's treatment. BBC News (November 1). Retrieved from: https://www.bbc.co.uk/news/technology-46054202

Lerchenmueller, M. J., Sorenson, O. and Jena, A. B. (2019). Research: How Women Undersell Their Work. Harvard Business Review (December 20). Retrieved from: https://hbr.org/2019/12/research-how-women-undersell-their-work

Leslie, S.-J., Cimpian, A., Meyer, M. and Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. Science, 347 (6219), 262 – 265.

Levin, S. (2017). As Google AI researcher accused of harassment, female data scientists speak of 'broken system'. The Guardian (December 22). Retrieved from: https://www.theguardian.com/technology/2017/dec/22/google-ai-researcher-sexual-harassment-female-data-scientists

Li, L., Yang, J., Jing, H., He, Q., Tong, H. and Chen, B. C. (2017). NEMO: Next career move prediction with contextual embedding. 26th International World Wide Web Conference 2017, 505–513.

LinkedIn. (n.d.) "LinkedIn". Retrieved from: https://www.linkedin.com/

Little, J. (2020). 'Everything has been pushed back': how Covid-19 is dampening tech's drive for gender parity. The Guardian (18 June). Retrieved from: https://www.theguardian.com/careers/2020/jun/18/everything-has-been-pushed-back-how-covid-19-is-dampening-techs-drive-for-gender-parity

Mahdawi, A. (2021). What a picture of Alexandria Ocasio-Cortez in a bikini tells us about the disturbing future of AI. The Guardian (February 3). Retrieved from: https://www.theguardian.com/commentisfree/2021/feb/03/what-a-picture-of-alexandria-ocasio-cortez-in-a-bikini-tells-us-about-the-disturbing-future-of-ai

Mantha, Y. and Hudson, S. (2018). Estimating the Gender Ratio of AI Researchers Around the World. Element AI (August 17). Retrieved from: https://medium.com/element-ai-research-lab/estimating-the-gender-ratio-of-ai-researchers-around-the-world-81d2b8dbe9c3

Margolis, J. and Fisher, A. (2002). Unlocking the Clubhouse: Women in Computing. Cambridge: MIT Press.

Maron, D. F. (2018). Science Career Ads Are Disproportionately Seen by Men. Scientific American. Retrieved from: https://www.scientificamerican.com/article/science-career-ads-are-disproportionately-seen-by-men/

Maurer, C. C. and Qureshi, I. (2019). Not just good for her: A temporal analysis of the dynamic relationship between representation of women and collective employee turnover. Organization Studies, 42 (1), 85 – 107.

Miltner, K. M. (2018). Girls Who Coded: Gender in Twentieth Century U.K. and U.S. Computing. Science, Technology, & Human Values, 44 (1), 161–176.

Misa, T. J. (2010). 'Gender Codes: Defining the Problem'. In: Gender Codes: Why Women Are Leaving Computing. In: IEEE Computer Society Press, Wiley, 3–23.

Mitchell, M., Baker, D., Moorosi, N., Denton, E., Hutchinson, B., Hanna, A., Gebru, T. and Morgenstern, J. (2020). Diversity and Inclusion Metrics in Subset Selection. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 117–123.

Murray, S. (2016). When female tech pioneers were the future. Financial Times (March 7), Special Report: Women in Business, 9–12. Retrieved from: https://www.ft.com/content/5a9d9cf2-d97b-11e5-a72f-1e7744c66818

Muzio, D. and Tomlinson, J. (2012). Editorial: Researching Gender, Inclusion and Diversity in Contemporary Professions and Professional Organizations. Gender, Work and Organization, 19 (5), 455–466.

Mylavarapu, S. (2016). The lack of women in tech is more than a pipeline problem. TechCrunch (May 10). Retrieved from: https://techcrunch.com/2016/05/10/the-lack-of-women-in-tech-is-more-than-a-pipeline-problem/

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

Oertelt-Prigione, S. (2020). *The impact of sex and gender in the COVID-19 pandemic.* European Commission Independent Expert Report (May).

Office for Artificial Intelligence (2019). £18.5 million to boost diversity in AI tech roles and innovation in online training for adults. Press release with Department for Digital, Culture, Media & Sport, Department for Education and Department for Business, Energy & Industrial Strategy. Retrieved from: https://www.gov.uk/government/news/185-million-to-boost-diversity-in-ai-tech-roles-and-innovation-in-online-training-for-adults

Oldenziel, R. (1999). *Making technology masculine: Men, women and modern machines in America.* Amsterdam: Amsterdam University Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown.

Ong, P. and Leung, I. (2016). "Tides are turning in favour of women in tech. Here's why." *Tech in Asia (August 25).*

OpenML (n.d.). OpenML. Retrieved from: https://www.openml.org/

Patel, K. (2020). 'Is AI Innately sexist?' *The World Today,* February and March. Chatham House. Retrieved from: https://www.chathamhouse.org/publications/twt/ai-innately-sexist

Paul, K. (2019). 'Disastrous' lack of diversity in AI industry perpetuates bias, study finds. *The Guardian (April 17).* Retrieved from: https://www.theguardian.com/technology/2019/apr/16/artificial-intelligence-lack-diversity-new-york-university-study

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S. and Niebles, J. C. (2019). The AI Index 2019 Annual Report. *AI Index Steering Committee, Human-Centered AI Institute.* Stanford University, Stanford, CA. Retrieved from: https://hai.stanford.edu/research/ai-index-2019

Posner, M. (2017). We can teach women to code, but that just creates another problem. *The Guardian (March 14).* Retrieved from: https://www.theguardian.com/technology/2017/mar/14/tech-women-code-workshops-developer-jobs

Prates, M., Avelar, P. and Lamb, L. C. (2019). *Assessing gender bias in machine translation: A case study with google translate.* Neural Computing and Applications, March.

Purtill, C. (2021). Hey, Alexa, Are You Sexist? *The New York Times (February 12).* Retrieved from: https://www.nytimes.com/2021/02/12/us/alexa-bots-female-voice.html

PwC. (2020). Women in Work 2020: The opportunities and challenges of the tech revolution. PwC.

Quirós, C. T., Morales, E. G., Pastor, R. R., Carmona, A. F., Ibáñez, M. S. and Herrera, U. M. (2018). *Women in the Digital Age.* Brussels: European Commission

Quotacom (n.d.). “Quotacom”. Retrieved from: https://www.quotacom.com/

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J. and Denton, E. (2020). Saving Face: Investigating the ethical concerns of facial recognition auditing. *AIES 2020: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 145–151.

Roca, T. (2019). Identifying AI talents among LinkedIn members: A machine learning approach. *Microsoft and LinkedIn Economic Graph.*

Schluter, N. (2018). The glass ceiling in NLP. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 22 (3),* 2793–2798.

Scott, A., Kapor Klein, F. and Onovakpuri, U. (2017). Tech Leavers Study. *Ford Foundation/Kapor Center for Social Impact.*

Scott, L. (2021). The Spark podcast: ‘Linda Scott and the Double-X Economy’. BBC Sounds (January 20). Retrieved from: https://www.bbc.co.uk/sounds/play/m000rdlr

Simonite, T. (2018). AI is the future – but where are the women? *WIRED (17 August).* Retrieved from: https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/

Sloane, M., Moss, E., Awomolo, O. and Forlano, L. (2020). Participation is not a design fix for machine learning. *Proceedings of the 37th International Conference on Machine Learning,* Vienna, Austria, PMLR 119.

Specia, M. (2019). Siri and Alexa Reinforce Gender Bias, U.N. Finds. *The New York Times (May 22).* Retrieved from: https://www.nytimes.com/2019/05/22/world/siri-alexa-ai-gender-bias.html

Stack Overflow (n.d.). 'Stack Overflow'. Retrieved from: https://stackoverflow.com/

Stathoulopoulos, K. and Mateos-Garcia, J. (2019). Gender Diversity in AI Research. *Nesta.*

Steed, R. and Caliskan, A. (2021). Image representations learned with unsupervised pre-training contain human-like biases. For: FAccT '21, March 3–10, 2021, Virtual Event, Canada. *ArXiv preprint.*

Strengers, Y., Qu, L. and Xu, Q. and Knibbe, J. (2020). Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation. In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Honolulu, HI.

Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J. and Schiebinger, L. (2019). Sex and gender analysis improves science and engineering. *Nature* 575, 137–146.

Tech Nation (2018). Diversity and inclusion in UK tech companies. Retrieved from: https://technation.io/insights/diversity-and-inclusion-in-uk-tech-companies/

Terrell, J., Kofink, A., Middleton, J., Rainear, C., Murphy-Hill, E., Parnin, C. and Stallings, J. (2017). Gender differences and bias in open source: pull request acceptance of women versus men. *PeerJ Computer Science,* 3:e111.

The Alan Turing Institute (2021). Frequently asked questions. The Alan Turing Institute. Retrieved from: https://www.turing.ac.uk/about-us/frequently-asked-questions

The Alan Turing Institute (n.d.). Women in Data Science and AI Hub. Retrieved from: https://www.turing.ac.uk/about-us/equality-diversity-and-inclusion/women-data-science-and-ai

The Alan Turing Institute. (n.d.). Women in data science and AI – Resources. Retrieved from: https://www.turing.ac.uk/about-us/equality-diversity-and-inclusion/women-data-science-and-ai/resources

The Royal Society. (2019). Dynamics of data science skills How can all sectors benefit.

Thompson, C. (2019). The Secret History of Women in Coding. *The New York Times Magazine (February 13).*

Tockey, D. and Ignatova, M. (2019). Gender Insights Report: How women find jobs differently. *LinkedIn Talent Solutions.*

Tulshyan, R. (2019). Do Your Diversity Efforts Reflect the Experiences of Women of Color? *Harvard Business Review.* Retrieved from: https://hbr.org/2019/07/do-your-diversity-efforts-reflect-the-experiences-of-women-of-color

UK AI Council (2021). *AI Roadmap.* Retrieved from: https://www.gov.uk/government/publications/ai-roadmap

UK Government (2020). “Guidance: Data Scientist”. Retrieved from: https://www.gov.uk/guidance/data-scientist

UNESCO (2020). Artificial Intelligence and Gender Equality: Key Findings of UNESCO’s Global Dialogue. *UNESCO.* Paris, France.

UN Women (2020). Spotlight on Gender, Covid-19 and the SDGs: Will the Pandemic Derail Hard-Won Progress on Gender Equality? *Spotlight on the SDGs*. Women Count/Covid-19 Response.

UN Women/Women Count (n.d.). “Women Count.” Retrieved from: https://data.unwomen.org/women-count

Vasilescu, B., Posnett, D., Ray, B., van den Brand, M. G. J., Serebrenik, A., Devanbu, P. and Filkov, V. (2015). Gender and Tenure Diversity in GitHub Teams. *CHI 2015*, Seoul, Korea. Gender & Technology, 3789 - 3798.

Vleugels, A. (2018). Want AI to be less biased? Cherish your female programmers. *TNW (January 11).* Retrieved from: https://thenextweb.com/artificial-intelligence/2018/01/11/want-ai-to-be-less-biased-cherish-your-female-programmers/

Wajcman, J. (1991). *Feminism Confronts Technology.* Cambridge: Polity Press.

Wajcman, J. (2010). Feminist theories of technology. *Cambridge Journal of Economics, 34 (1),* 143–152.

Wajcman, J., Young, E. and FitzMaurice, A. (2020). The Digital Revolution: Implications for Gender Equality and Women's Rights 25 Years after Beijing. *Discussion Paper No. 36 (August),* UN Women.

West, M., Kraut, R. and Chew, H. E. (2019). *I'd blush if I could: closing gender divides in digital skills through education.* UNESCO Equals, 306.

West, S., Whittaker, M. and Crawford, K. (2019). *Discriminating Systems: Gender, Race and Power in AI.* New York: AI Now Institute, New York University

World Economic Forum (2018). Global Gender Gap Report 2018. Retrieved from: https://www.weforum.org/reports/the-global-gender-gap-report-2018

World Economic Forum (2020a). Global Gender Gap Report 2020. Retrieved from: https://www.weforum.org/reports/gender-gap-2020-report-100-years-pay-equality

World Economic Forum (2020b). The Future of Gender Parity: A Labour Market Shift. Retrieved from: https://reports.weforum.org/global-gender-gap-report-2020/the-future-of-gender-parity/?doing_wp_cron=1612906910.1824119091033935546875

Wynn, A. T. and Correll, S. J. (2017). Gendered Perceptions of Cultural and Skill Alignment in Technology Companies. *Social Sciences*, *6* (2), 45.

Wynn, A. T. and Correll, S. J. (2018). Puncturing the pipeline: Do technology companies alienate women in recruiting sessions? *Social Studies of Science*, 48 (1), 149 – 164.

Yates, K. (2020). Why do we gender AI? Voice tech firms move to be more inclusive. *The Guardian (January 11).* Retrieved from: https://www.theguardian.com/technology/2020/jan/11/why-do-we-gender-ai-voice-tech-firms-move-to-be-more-inclusive

Young, E. (2020). "Technology and inequality in the era of pandemic: data, power and unrest". *The Alan Turing Institute (August 11).* Retrieved from: https://www.turing.ac.uk/blog/technology-and-inequality-era-pandemic-data-power-and-unrest

Zacharia, Z. C., Hovardas, T., Xenofontos, N., Pavlou, I. and Irakleous, M. (2020). Education and employment of women in science, technology and the digital economy, including AI and its influence on gender equality. (April 15). *European Parliament. Policy Department for Citizens' Rights and Constitutional Affairs (IPOL).* Retrieved from: https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2020)651042

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J.C., Sellitto, M., Shoham, Y., Clark, J. and Perrault, R. (2021). "The AI Index 2021 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V. and Chang, K. W. (2019). *Gender Bias in Contextualized Word Embeddings*. Retrieved from: http://arxiv.org/abs/1904.03310

Zmigrod, R., Mielke, S. J., Wallach, H. and Cotterell, R. (2019). Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* July.

Zou, J. and Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature,* 559, 324–326.