Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–March: 1–11

DOI: 10.1177/20539517231221780
journals.sagepub.com/home/bds

Sage

Tech workers' perspectives on ethical issues in AI development: Foregrounding feminist approaches

Jude Browne[1], Eleanor Drage[2] and Kerry McInerney[2]

Abstract

While tech workers are essential stakeholders in ethical artificial intelligence (AI) development and deployment, they are rarely consulted about their understanding of the development of ethical AI. In light of this, we present the findings of our 2020 to 2021 empirical research study in which we collected data from tech workers in a major AI company to better understand what they consider to be the most pressing ethical issues when developing AI-powered products. While there is a nascent body of literature that examines how AI ethics principles are operationalised on the ground, this study differs in that we explicitly draw on feminist insights to inform our analysis, and have put a particular focus on allowing the voices and narratives of tech workers to lead the work forward. Our study generated three main findings: first, the term 'bias' creates real confusion among tech workers, meaning that the term is unable to do the ethical work it is intended to do; second, tech workers do not necessarily see a relationship between diversity, equality and inclusion (DEI) agendas and AI development, undermining AI ethics initiatives; and third, tech workers were particularly concerned about the monitoring and maintenance of unwieldy 'legacy systems' that generated serious challenges to creating and deploying new and more ethical AI products. This study thus creates a 'thicker' and more nuanced picture of tech workers' perspectives on the ethical issues that arise when developing and maintaining AI systems, while simultaneously demonstrating the utility of feminist approaches in the field of AI ethics.

Keywords

Artificial intelligence, AI ethics, feminism, big tech, STS, industry

Introduction

When seeking to ensure that artificial intelligence (AI) is developed ethically in today's fast-moving and demanding economic environment, we might assume that it is crucial to listen to tech workers' needs and concerns. However, as research has shown, employees are rarely consulted when corporations develop AI ethics strategies, largely due to the assumption that ethicists are at an epistemological advantage when it comes to 'thinking ethically' (Hagendorff, 2023). This exclusion is particularly apparent in the case of tech workers who do not come from engineering or STEM backgrounds, such as lawyers, marketeers or human resources (HR) professionals.[1] These employees are often disproportionately impacted by new technologies or on the front line of product deployment, yet their opinions and ethical feedback are rarely sought or incorporated in the product development and design process.

In this article, we aim to combat this lack of insight into tech workers' perspectives on AI ethics through the findings of our empirical study at a major technology multinational. We conducted in-depth qualitative interviews with over 60 employees in a range of fields – including AI, data science, legal services and HR – at Tech Company X (TCX), a company headquartered in the Global North that is one of the foremost AI patent holders in the world and employs over 100,000 staff globally. Our research differs from the small number of existing empirical studies (Khan et al., 2023; Powell et al., 2022; Ustek-Spilda et al., 2019; Vakkuri et al., 2020) in that we explicitly draw on feminist insights and have put a particular focus on allowing the voices and narratives of tech workers to lead the work forward. We follow the methodological approach of

[1]Centre for Gender Studies, University of Cambridge, Cambridge, UK
[2]Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

Corresponding author:
Kerry McInerney, Leverhulme Centre for the Future of Intelligence, University of Cambridge, 16 Mill Lane, CB2 1SB, Cambridge, UK.
Email: kam83@cam.ac.uk

feminist and critical race scholar Sara Ahmed (2012) by permitting interviewees – in our case, tech workers – to tell their own stories about AI ethics, and we draw on feminist political theory and feminist science and technology studies to interpret them. Our feminist approach and our focus on storytelling complicate the assumption that corporations are monolithic entities with singular understandings of AI ethics, and instead emphasise the importance of employees' disparate ethical views and understandings of what AI can and should do. By exploring the perspectives of tech workers who possess different forms of domain knowledge but support the operation of the same organisation, we show how, even within a single company with a globally recognised brand and a highly developed internal corporate culture, the meaning and significance of AI ethics substantially shifts between individuals and groups. This has serious ramifications for how the development of ethical AI is perceived and understood across the whole sector.

In allowing tech workers to tell their own stories about AI, three main themes or concerns rose to the surface: what the term 'bias' meant to tech workers and how they believed the concept fits into their work; their perceptions of how the development of AI relates to diversity, equality and inclusivity (DEI) issues and corporate diversity initiatives; and the challenges for developing ethical AI posed by integrating new AI systems with existing 'legacy systems' and successfully maintaining AI systems over time.

The pressure points we identify from tech workers' stories at different steps in the AI development and deployment pipeline ultimately demonstrate the limitations of approaching the development of ethical AI as a one-way flow of information from AI ethicists and philosophers to tech workers in industry contexts. While we are sceptical of the turn toward self-regulation as the primary response to the ethical issues generated by AI and other data-driven technologies (Kak and West, 2023), we draw on feminist concepts such as Donna Haraway's 'situated knowledges' to show how AI ethics must remain grounded in the contexts where technologies are developed and deployed. If not, AI ethics strategies risk becoming little more than 'ethics washing' or high-level principles that tech workers struggle to embed into industry practice. Thus, the findings from our study provide key insights for three main groups. First, they can assist industry and tech workers attempting to develop ethical AI at the company level, generating greater insight into how individuals are thinking about AI and how ideas about AI percolate within and across organisations. Second, they can improve AI ethicists' understanding of how tech workers on the ground are interpreting and implementing AI ethics, showing them what languages and ideas are currently working as intended and where there is room for improvement. Third, our findings can inform policymakers and experts in AI governance about tech workers' attitudes toward the development of ethical AI at the micro-level, and these findings may therefore influence their decision-making at the macro-level of AI governance.

Understanding AI ethics in practice

Companies, international organisations, governments, industry, public sector organisations, academic institutions and civil society groups have converged on the concept of 'AI ethics' as a way of guiding AI development and deployment. In the corporate sector, notable contributions to AI ethics include Google's Responsible AI, Microsoft's Responsible AI, Rolls-Royce's Aletheia Framework (2021), BMW's Seven Principles for AI (2020), PwC's Practical Guide to Responsible AI (2019) and Digital Catapult's Ethics Framework (2022). Meanwhile, over 60 countries worldwide have developed their own AI ethics strategies (OECD, 2021), encouraging the development of just, fair, transparent and accountable AI systems, while simultaneously addressing associated practices such as data collection and management. At the international level, a multitude of AI ethics frameworks have been produced, including NATO's Principles of Responsible AI (2021) and the European Parliament's Governance Framework for Algorithmic Accountability and Transparency (2019). While different AI ethics projects contain their own distinct political agenda and are aimed at different stakeholder audiences, there are some identifiable shared concerns. As Jobin et al. (2019) map out in their study of 84 international ethics frameworks, there are 11 common key themes.[2] Of particular interest to this study is the second most prominent theme: '*justice, fairness and equity*', which Jobin et al. (2019) suggest 'is mainly expressed in terms of fairness, and of prevention, monitoring or mitigation of unwanted bias and discrimination' (p. 394). We suggest that by listening to tech worker's views and stories, a more complex picture emerges of how these generic ethical terms are insufficient for effective change.

Existing empirical research already highlights some of the difficulties that emerge when attempting to implement AI ethics frameworks in industry contexts. Common problems include outsourcing AI ethics to high-level committees; appointing a single individual to be in charge of ethical practice (rather than involving the whole development team); and assuming that ethical approaches will be implemented without systematic guidance, review and oversight (Khan et al., 2023; Vakkuri et al., 2020). Vakkuri et al. (2020) surveyed 249 practitioners at 211 software companies (106 of which were developing AI) about key issues and principles in AI ethics in order to better understand how tech workers were approaching ethical questions on the ground. They conclude that 'the data we collected point to AI ethics implementation still being in its infancy'

(Vakkuri et al., 2020: 54) and highlight that engineers' perspectives must be included in AI ethics discourse and policy in order for it to be actionable: 'using these guidelines requires additional work from your organization because they do not come in the form of an off-the-shelf method. You need to first make them more practical for your developers, project teams, and product owners and customers' (Vakkuri et al., 2020: 54). Meanwhile, an international survey of 99 tech workers and lawmakers by Khan et al. (2023) explored what these practitioners thought were the most important AI ethics principles and the main challenges they faced in their implementation. The most commonly stated barrier to the implementation of AI ethics principles (as cited by 81.8% of interview participants) was a 'lack of ethical knowledge' that made it difficult to expand and scale AI ethics principles in industry contexts (Khan et al., 2023: 4).

This has led some critics to suggest that high-level AI ethics frameworks and strategies function primarily as public relations exercises or as a form of 'ethics washing' (Burt, 2020; Haas and Gießler, 2020; Hao, 2019; Kerr et al., 2020; Metzinger, 2019; Munn, 2022; Rességuier and Rodrigues, 2020). For example, Rességuier and Rodrigues (2020: 2) posit that AI ethics frameworks have come to stand in for the effective regulation of the AI sector, providing few legal or political protections for those who are most likely to be adversely affected. Others have argued that the field of AI ethics has created a new 'economy of virtue' that produces 'ethics as a product' (Phan et al., 2022), aptly demonstrating what Meredith Whittaker (2021) calls the steep cost of corporate capture in AI ethics. The framing of AI ethics as 'ethics washing' or 'ethics theatre' corresponds to feminist analyses of DEI policies that stand in for the systemic change that they call for: that is to say, they are 'non-performatives' (Ahmed, 2012: 6). Certainly, AI ethics strategies frequently use terms such as 'responsible AI' and 'trustworthy AI' without determining what these words mean or how they are to be implemented and measured (Jobin et al., 2019). After all, many AI ethics frameworks operate on the assumption that tech workers are in agreement over key aspects of AI ethics, such as what AI is, what constitutes a biased system, which systems require ethical attention, how the regulatory environment affects them, how AI ethics relates to the various cultures of the AI sector and how different populations are affected by AI systems (see e.g. Rolls-Royce's Aletheia Framework (2021), BMW's Seven Principles for AI (2020), PwC's Practical Guide to Responsible AI (2019) and Digital Catapult's Ethics Framework (2022)).

The existing literature and the critical response to AI ethics demonstrate the value of gathering empirical data on the kind of AI ethics issues tech workers lack clarity on, and what their major concerns are when it comes to the development of ethical AI. However, while these broad surveys provide a useful starting point for understanding what AI ethics looks like and means 'on the ground', they leave significant scope for more in-depth, qualitative analyses of tech workers' attitudes toward and distinct struggles with AI ethics issues. Finally, these studies survey a range of different firms and organisations. While this provides a helpful litmus test of the general mood of the AI industry as a whole, we believe that an in-depth study of one leading site of AI development and deployment – a major technology company – can illuminate the kind of barriers individual organisations face when seeking to develop and deploy AI ethically. This more focused study of TCX allows us to investigate how ideas about AI ethics move across different organisational domains such as engineering, data science, HR and legal services.

Methodology

In this article, we draw on the insights of feminist theory and qualitative methods to understand what AI ethics is 'doing' on the ground in a major technology multinational. Central to our study is the feminist concept of situated knowledge. Coined by feminist philosopher of science Donna Haraway (1988), it emphasises that all forms of knowledge are shaped by the particular conditions in which they are produced. Hence, while some AI ethicists emphasise universal truths of moral reasoning and attempt to graft these principles onto industry contexts (Zavalishina and Polonski, 2017), we argue instead that it is crucial to explore the local and situated perspectives of those who are 'doing' the work of creating ethical AI. Additionally, feminist and anti-racist scholars such as Benjamin (2019a, 2019b), Buolamwini and Gebru (2018), Noble (2018), Crawford (2021), D'Ignazio and Klein (2020) and Hampton (2021) show how considerations of AI as 'ethical' cannot be divorced from the people and institutions that develop and deploy them.

Our industry partner in this project, TCX, wishes to remain anonymous and accordingly, we have been extremely careful not to include any information that might disclose TCX's identity or those employees who took part in our research. In addition to removing the name of the company and any details about interviewees' identity, we have also excluded geographical locations and references to specific AI products. In the spirit of ethical progress, TCX spent a great deal of time working with us on this project. One of our team, the Principal Investigator, established a professional relationship with TCX over several years, which resulted in an invitation to partner on this project. TCX entered the partnership, built on a detailed non-disclosure agreement with the research team to protect anonymity, with a genuine curiosity and willingness to be observed 'from the inside'. A senior manager of TCX was appointed as the team's point-person,

responsible for making the project visible to employees and helping us to seek out interviewees from across its workforce. Senior staff also gave us a detailed presentation on the structure and objectives of TCX, and helped us understand how AI is used, designed, developed, tested and marketed.

To initiate the study and explain our methodological approach, two members of our team gave an online presentation on the project which was open to all employees and recorded for those who could not attend. We were aware that the presentation might deter some interviewees from agreeing to participate. Nevertheless, we thought it was important to emphasise our genuine effort to explore tech worker attitudes towards the ethical issues arising in AI development and deployment. As evidenced by our data, employees were extremely open and candid in their responses, and expressed a very wide range of perspectives.

Initially facilitated by our point-person, an invitation to engage with our project was promoted across TCX and over the course of 2020 to 2021 we interviewed 63 TCX employees who volunteered to be part of the project.[3] Because interviews were voluntary and based on generic invitation emails, and because of the snowball effect generated by employees' peer-to-peer recommendations, participants were necessarily self-selecting. Nonetheless, they worked in a wide range of fields, including AI research and design, data science, legal services and HR, and their seniority ranged from new intake to senior management. We aimed to collect as diverse a sample of the organisation as possible based on demographic information (age, self-defined gender, ethnicity and registered disabilities). We developed our final aide-mémoire based on the emergent themes in an initial pilot of 10 interviewees. In-depth interviews were conducted by the research team in private and anonymity was offered to all participants. Interviews tended to range from 30 min to 1 h depending on how much time participants required to tell their stories and give their views.

Following feminist scholar Sara Ahmed's (2012) methodological approach in *On Being Included*, our study explored the situated knowledge of tech workers through open-ended conversations. We recognise that with the in-depth discussion-based interview methods we adopted, we were only able to reach a small proportion of the total number of TCX employees. We were also conscious of Ahmed's warning that too much research with institutions 'is founded on what institutions want found' (2012: 10), and so we chose to use Ahmed's particular qualitative methods in order to allow employees at TCX to develop and tell their own stories about AI, and to let their distinct concerns and views rise to the surface. As Ahmed (2012) notes, the possibilities of 'what can be found' are limited by data collection methods (p. 10). Rather than relying on the narrower categories of surveys (that often predetermine interviewees' answers based on the questions set) or the more rigid interview Q&A structures of questionnaire-based research (that tend to fixate the researcher on collecting strictly comparative answers), we, like Ahmed, used our aide-mémoire to begin open-ended conversations with our subjects about their experiences of working with AI. We encouraged practitioners to explore their own feelings and expectations about the technologies they were helping to produce, test, use and market. As feminist scholar Young (2000) argues, by allowing people to tell stories of their experiences, we are more likely to hear the themes and perspectives that are important to interviewees, which might otherwise be overlooked as irrelevant or too trivial to be captured by more structured methods. Furthermore, as feminist scholars Jugov and Ypi (2019) argue, because so many of us are unaware of our relationship to structural injustice and discriminating societal phenomena, there needs to be more methodological focus on hearing each other's experiences in order to build a clearer picture of the conditions in which people make decisions. This is what feminist scholar Haslanger (2015: 15) calls the 'choice architecture' of our daily lives. Only when we can determine something of tech workers' choice architecture when building, marketing and deploying AI can we begin to respond to the challenges that they face in developing AI ethically.

However, while we allowed employees to tell their own stories, we did not approach these stories uncritically, nor did we assume that they contained the sole 'truth' about AI and the challenges tech workers faced in developing AI ethically. Following feminist political theorist Banu Bargu (2013), we took 'neither the dominant narratives of power nor the narratives subjugated by them at their word' (p. 806). Instead, we approached the stories told by employees generously and sensitively, while simultaneously recognising how their experiences may obscure or occlude the real harms potentially generated by their products. This refractive practice between narrative and theory allowed us to fully honour the narratives of employees at TCX while also recognising how their stories play into the wider gendered and racialised power dynamics that constitute the AI industry.

Our original intention had been to travel to a range of different TCX locations and conduct ethnographic research with our subjects over the course of their usual working days. However, due to the national and international lockdowns prompted by the COVID-19 pandemic, we resorted to Zoom interviews. While this prevented observations and interactions with interviewees in situ, there were nevertheless a number of positive effects. In speaking to us from home, participants had more privacy from their colleagues when discussing their views and perspectives. There were also far fewer logistical challenges in organising face-to-face interviews, which we believe gave rise to a higher number of interviewees agreeing to join us for a substantial discussion (Gray et al., 2020: 1292).

We recorded the interviews in 2021, saved them locally to our computers and used Otter AI to begin the transcription process. The generated transcripts were then checked and amended against the original audio by our research team. The documents were read multiple times by our research team so that they became familiar with the texts and key themes raised. NVivo was used for coding and to assist with data analysis. A great many issues were discussed surrounding AI at TCX; however, for the purposes of this article we have concentrated on three central themes that emerged from the data: understandings of bias, the relationship between DEI and AI development, and the practical challenges to developing ethical AI that arise from integrating new AI systems into older technological infrastructures.

Results

Bias

We asked interviewees what bias meant to them and how it related to their work, which raised a wide variety of issues. Among our participants, bias was primarily associated with flawed training data. However, a wide range of other concerns emerged, including algorithmic drift (whereby the relationship between data and predictive models shifts over time), the poor management of data, abuse by ill-intentioned users, and the commercial de-prioritisation of marginalised users and consumers. A significant number of practitioners (10 out of 63) were unsure where and how bias might appear.

Not only were tech workers confused about where bias comes from, they also often did not see their own work as having any relationship to AI-generated bias at all:

> I'm fortunate, in a sense, in that my models don't tend to make any predictions about people or things like that … So, we're just trying to see if we can build a bot that's capable of playing a game with the user to be able to solve a problem, right? So, that's all just generated conversational data. So, it doesn't involve any historical data. So thankfully, we don't have to deal with those problems. (#39)

This example demonstrates how some employees are not taught to question the basic assumptions made about the psychological, social and physical capabilities and reactions of the human 'player', and how such an uncritical approach might have direct or indirect ethical consequences for AI and robotics design. They, like several others,[4] viewed some data and systems as supra-ethical – beyond the scope of ethical attention. Several of these tech workers stated that bias was only a potential issue if data was directly 'about people'[5] and if not then, '[i]t's just pure, pure data … that's feeding the models' (#54). For example, one interviewee stated that their work did not evoke any ethical issues because they were not using 'personal data', despite explaining that the training datasets for their work came from Wikipedia entries:[6]

> I felt like it wasn't personal to the user or the data subjects … Obviously you can't account for what's happening in that machine learning process because it was unsupervised learning, but even so, I felt like there weren't enough indicators within the [Wikipedia] text to indicate any form of gender, or any other personal group. (#35)

This confusion stemmed, in part, from disciplinary differences over the meaning of the term 'bias'. Not all usages of this term had a negative connotation of 'unfairness' for tech workers. While many tech workers were personally invested in eradicating injustices such as racism and sexism, some understood bias in its mathematical sense, using the term to indicate 'significant trends in the data'[7]:

> Bias can be good, right! In my field, in terms of trying to find a threat … you want to … bias towards finding threats rather than being fair … You know, we are by definition discriminating in the data. (#41)

> [B]ias is … like the intercept from the y axis. (#62)

These responses demonstrate that interpretations of bias are strongly contingent on an employee's disciplinary specialism, an issue which is rarely raised in public and policy discourse on AI ethics. Additionally, the fact that the interpretation of bias in AI ethics discourse is not as familiar to engineers is likely to reflect how computer science education and tech companies put less emphasis on bias defined as unfairness than they do the kinds of bias that engineers alluded to: #41 is referring to bias as the attempt to direct a model towards identifying possible threats, and #62 is discussing bias as a feature of the model that helps it learn and improve performance. Both responses show a primary concern for product performance: in the first case by seeking to eradicate threats to the product, and in the second by focusing on the improvement of functionality.

Furthermore, both give insight into how engineers perceive AI ethics attempts to debias systems. By contrasting their own definition of bias to 'being fair', #41 gives insight into the common engineering perspective that the eradication of bias as part of AI ethics initiatives means making the system 'equal' or 'neutral'. However, implicit in his response is the fact that AI models are always 'skewed' in their visions of the world – because you have to bias 'towards' a given output. This points towards the impossibility of rendering a system neutral through debiasing techniques. As we explore in this article, it is crucial that AI ethics discourse learns from engineers' understanding of

AI systems as always limited, in that they fundamentally take a partial, biased perspective.

Other respondents also signalled that that data could not be debiased or neutralised:[8]

> There is the potential to lose the humanity of people we're categorising with AI … You know, the trouble with AI I think is that it can only be so fine-grained … it has to put people in a box maybe unfairly, you know, to do anything, because that's all it's capable of doing. (#6)

Again, we see an acknowledgement that systems offer low-definition approximations of the world which can 'only be so fine-grained' and have 'to put people in a box', which necessarily engenders a loss of the complexity of human experience. These insights show the potential for engineers to shift AI ethics work away from the impossible task of rendering systems neutral and towards ensuring that the orientation taken by AI is ethical and serves the interests of, for example, minoritised groups.

Relationship between AI development and DEI

Despite high-level corporate messaging at TCX that emphasises the importance of institutional diversity, most tech workers did not make a direct connection between DEI and AI development. One-third of our interviewees (21, or 33%) said that they thought diversity was an important priority for TCX, or that it was a positive force in their teams and in TCX more broadly.[9] However, only a far more limited number of interviewees (10, or 16%) connected DEI initiatives to the development of AI, and their views on the utility of AI ethics in the actual AI design and development process varied significantly across domains and between individuals.[10] While one AI practitioner was not at all convinced by the potential of diversity for product development (#3), others felt that TCX was not creating strong enough links between its DEI programmes and AI development.[11] For example, interviewee #19, speaking about the company's programme to increase the representation of ethnic minorities, said:

> So [the diversity initiative] is an HR program … because of that it doesn't really touch the technologies that we're developing, it relates more to the people that work for us and what we've got in place to help them to progress personally. (#19)

While many of the employees we spoke to were passionate about addressing sexism and racism within the AI industry, some of our interviewees believed that attempts to diversify TCX's workforce were in fierce competition with its drive to hire more AI engineers.[12] Interviewee #29 noted, for example, that the need for a digitally skilled workforce might eclipse TCX's DEI ambitions by introducing more of the dominant STEM demographic into the company (often white and male) and incentivising the company to prioritise reskilling existing digital workers as opposed to bringing new, more diverse talent into TCX. This begs the question: which is a higher business priority – diversity or AI literacy? (#29). This competing set of commercial priorities was exacerbated by the high demand for tech talent when we conducted our interviews. As one high-level HR employee explained, 'We're literally calling it a war for talent at the moment', arguing that TCX could not afford to prioritise diversity over retaining and attracting highly desirable AI engineers (#46).

Managing legacy systems and AI over time

When asked about the main challenges they faced in developing AI ethically, a number of TCX tech employees (10, or 16%)[13] focused on the negative impact of 'legacy systems', or existing technological or institutional systems, on the development and deployment of new AI systems:

> We've developed things far beyond what we should have done; we should have replaced them earlier. We don't switch things off. We don't have an exit plan for any of the things that we build … As a consequence, when we develop new things and new products, we find that we have to rely on the old systems to deliver the new products. (#48)

In particular, several[14] employees (5, or 8%) thought that organisational 'sunk costs' prevented them from rebuilding systems that were no longer adequately functional:

> I do also think we've got a mountain to climb to unpick the organic mess, frankly, that we've created in the way that we've joined systems together … Basically, there's so many things kind of just bolted together. And it undermines the value of the data … as I'm sure is the same for many major companies. (#49)

> We've missed a trick here. Actually, it's about sort of fixing the front end, rather than patching 'bolt ons' onto the back end. (#60)

Others emphasised the challenges of managing legacy systems, and believed that this could translate into a lack of proper protocol around how AI is monitored over time.[15] This was a particular ethical concern for our interviewees as AI systems require constant attention and adjustments throughout their lifecycle in order to remain safe, fair and accurate:

> [What is important] is making sure that you've got the ability to monitor … over time as the data changes or as

> the machine learning model evolves … and that ultimately we switch things off if and when they're not needed anymore … How are these things going to be maintained over a long lifespan? Because organisations are typically quite bad at keeping on top of legacy. At some point AI will be legacy. (#25)

> It emerged that one of the data feeds the model was expecting to receive had been turned off and had been … broken for some time. And that was causing the model to make misdiagnoses. When it was caught, it was like, how did we not see this sooner? (#39)

Discussion

In the following section, we explore the three major themes of our research findings: AI-generated bias; DEI as a central feature of ethical AI; and the practical challenges for developing ethical AI now and into the future.

AI-generated harms: From bias to partiality

AI ethics policy and discourse foreground the problem of unfair bias in AI systems (Institute for the Future and Omidyar Network, 2018: 43; Rolls-Royce, 2021: 6; UNESCO, 2021) and prioritise the development of bias-mitigation strategies.[16] However, our study suggests that tech employees do not have a shared definition of 'bias' (often not using the term in the same way as AI ethics initiatives do), and also differ in their views about which systems and data can become biased. And yet, the way the term bias is interpreted and employed on the ground is not frequently the subject of interrogation in AI ethics, despite a range of scholarship which suggests that bias is an improper term for unfair or discriminatory aspects of AI systems, and that the eradication or mitigation of bias is not always an effective approach to reducing AI harms (Drage and Frabetti, 2023; Gebru, 2019; Prescod-Weinstein, 2018).

In our data, engineers often took a selective view of which kind of data and systems could become biased. Recall, for example, the engineer who felt that there was no 'form of gender, or any other personal group' represented in the Wikipedia entries used to train data (#35). However, we suggest that all entries on Wikipedia are profoundly shaped by private interests, ranging from individuals to interest groups and other organisations, including those paid to write entries (Osman, 2014) who have a vested interest in promoting certain information and perspectives on Wikipedia articles. On the specific question of gender, women are far less likely to edit Wikipedia or have entries written about them, meaning that data about women is less likely to be represented in a dataset drawn from Wikipedia (Tripodi, 2021: 2–3; see also Lemieux et al., 2023). Therefore, our data demonstrates that tech workers could benefit from a greater understanding of how structural inequality shapes the data they use.

While 'bias' may not be the appropriate term with which to communicate harms deriving from AI to engineers, we respect tech workers' interpretation of bias as a mathematical concept meaning 'patterns of significance', and so do not seek to correct the way engineers interpret this polysemous term. Instead, given that in our data several engineers suggest that systems are inevitably biased and cannot be impartial, we need a term that expresses how partiality is often the cause of AI's harms. Inspired by the theorising of Haraway (1988) and Amoore (2020), we encourage both AI ethicists and tech workers to think about 'AI partiality' – the partial perspective taken by every Machine Learning model. AI partiality derives from feminist theories of situated knowledge that claim that knowledge is context-specific rather than universal. The feminist concept of situated knowledge suggests the impossibility of 'pure' or 'objective' data. Rather, data should be understood as always laden with the uneven experiences of the unequal contexts in which it emerges and can only ever provide a partial view of any given phenomenon (Amoore, 2020; D'Ignazio and Klein, 2020; Haraway, 1988).

We argue that AI partiality can remind tech workers as well as users that AI (and the data it is built on) is not, and never can be, objective. While technologists would probably be aware that technological systems, data models and 'ontologies' take only a *partial view* of the world as they occupy a 'reduced vector space' (Chan et al., 2010), they might be less likely to consider how AI systems might exclude other important 'viewpoints' or privilege certain users over others. In this way, we suggest that the limitations of an AI system should become an emphasised feature of its definition which is likely to be much more productive for AI ethics than advocating for 'bias eradication' and other attempts to return systems to a state of neutrality. Shifting tech workers' perceptions of what AI is capable of – only a partial perspective rather than a definitive, objective perspective – can help combat impossible efforts to make products neutral and bias-free, which ultimately do little to make systems less harmful for consumers. Because discrimination often operates under a 'veneer of objectivity' (Benjamin, 2019b: 1), the concept of AI partiality also invites tech workers to notice how the situated judgement of an AI system might result in outcomes that disadvantage gendered, racialised and other minoritised populations.

DEI and AI development

As a myriad of AI ethics frameworks have suggested (Jobin et al., 2019), a product team's diversity (or lack thereof) affects AI design, development and deployment. The demographics of the AI sector are still, however, notoriously homogeneous, to the extent that West et al. (2019: 3) note

that AI is suffering from a 'diversity crisis'. In the US context, for example, only 26% of AI workers are women (Wodeki, 2021), and in the United Kingdom only 20%, leading Young et al. (2021) to argue that the field is characterised by structural inequality. In an attempt to correct the lack of women in AI, AI ethics initiatives emphasise the need for increased diversity and representation in design teams. For example, IBM's Coercive Control Resistant Design principles insist that 'having a diverse design team broadens the understanding of user habits, enabling greater exploration of use cases, both the positive and the negative' (Nuttall, n.d.: 2). We found in our data however that while DEI was an important priority for TCX, most of our interviewees did not see any relationship between attempts to 'diversify' the workforce and the development of AI ethics. Indeed, some interviewees wondered if TCX's desire for a digitally skilled workforce would undermine its ambitious DEI targets because of the limited diversity of the AI talent pool. While TCX has put a great deal of effort and resources into its DEI policies, we find the work of feminist and critical race scholars such as Ahmed (2012) and Walcott (2019) informative here as they have problematised the way that 'diversity' language can fail to perform the structural shifts necessary to achieve meaningful change (Ahmed, 2012: 117). In these cases 'diversity' loses its transformative intent and can instead function as a discourse of 'benign variation' from the norm (Mohanty, 1989: 18), while 'inclusion' incorporates marginalised people into existing structures that have not been designed with their inclusion in mind (Ahmed, 2012: 163). This may be why some interviewees saw DEI initiatives as a potential barrier to hiring AI engineers. When diversity is seen as a tick-box exercise, rather than as an ongoing process, tech workers may struggle to understand why diversity matters in the first place and its central role in creating ethical AI. For example, in our dataset only a small subset of tech workers believed that a team's diverse perspectives impacted the character of technology development. In a similar way to those tech workers at TCX who thought data not explicitly 'about people' was inherently bias-free, some of our interviewees did not see the products they made as reflective of themselves or the teams that made them.

We suggest that DEI initiatives and AI ethics goals need to be brought into direct conversation with one another so that companies, policymakers and AI ethicists can draw on different methodologies and approaches to gender, race and other social characteristics in such a way that they are not understood as merely 'personal attributes' or 'identity categories', but rather as a set of *relations* that shape how institutions and the individuals within them operate (Ahmed, 2012: 13; Drage and Mackereth, 2022). By approaching the problem of diversity not solely as an issue of the underrepresentation of specific individuals but also as an occlusion of different forms of situated knowledge, governments and other institutions can move beyond the narrative of the 'diversity crisis' in AI talent and start to better contextualise the development, function and social impact of AI. Such an approach may also involve ultimately moving beyond the language of diversity itself. For example, Costanza-Chock (2020: 6) pushes not just for improved diversity in design teams but for a fully realised vision of 'design justice' that centres the voices of those who are usually marginalised in technology design and development processes. Similarly, Browne (2023, 2024) argues for new regulatory models for AI that include diverse lay-members of the public with different sets of interests.

Legacy systems, data detachment and the future of AI ethics

Our data suggests that legacy systems are one of the principal obstacles to developing AI ethics. Legacy systems and insufficient 'exit plans' for AI are rarely addressed by AI ethics strategies, even though they were a common concern across our dataset. Indeed, the Frankenstein's monster systems created by grafting together old and new systems frequently failed to address the root problems that the new algorithms were meant to solve. If legacy systems determine how new AI-powered tools function, then AI ethics initiatives must also account for legacy systems and how AI products will be deployed in legacy contexts.

The ethical oversight of legacy systems might entail, for example, an investigation of potential ethical issues associated with the dispersal or reconstitution of data within an organisation or sector. This is particularly challenging when data becomes detached from its original context over time. Temporal separation – the time that passes between the collection and use of data – can result in engineers, down the line, being less familiar with the context in which the data was gathered. Meanwhile, spatial separation – when data is moved across domains – can result in a loss of contextual information. Feminist scholars D'Ignazio and Klein (2020) use the term 'strangers in the dataset' to describe data scientists who are estranged from the context and provenance of the data they are working with (p. 130). They argue that the levels of remove between data scientists and data collection and maintenance – whether temporal or spatial – increase the risk of creating AI that is unrepresentative of the communities and contexts from which the data was collected. This sort of data detachment also increases the likelihood that data scientists will not be aware that a given dataset might not be representative of other excluded communities.

These concerns highlight the fact that AI ethics is necessarily a long-term and complex endeavour. It cannot simply be achieved through one-off 'bias checks' in the development process. Maintenance is a

neglected area of AI study; as Pendergrast and Pendergrast (2021, np.) argue, 'to think about AI through the lens of maintenance practices is one way to acknowledge the long life of technological systems and their impacts on people and the environment'. Feminist studies of political economy have long demonstrated the importance of maintenance work, and how this work tends to be both undervalued and yet also critical for societal functioning (see, e.g. Baraitser, 2017; Boyer, 2015; Fraser, 2016; Peterson, 2003). The same, we argue, applies to technical systems: maintenance work such as debugging was seen by some of our interviewees as boring and undesirable work. Nonetheless, they emphasised how monitoring and calibrating AI systems to ensure their fair and consistent performance over time is vital to the long-term practice of AI ethics. In light of these observations, we suggest that the practice of maintenance be prioritised in AI ethics strategies going forward.

Conclusions

This study has demonstrated the importance of listening to tech workers' own stories of the ethical issues arising from AI development in order to better understand what AI ethics means in local contexts. Our focus on a major tech company headquartered in the Global North – while not indicative of the entire tech industry – allows us to ground often-abstracted AI ethics discourse and concepts in the material realities of industry practice. This is critical for ensuring that AI ethics work meaningfully addresses employees' concerns, while simultaneously keeping in sight the wider power relations that structure the AI industry and lead to some ethical concerns being given greater attention than others. To ensure that AI ethics does not become a mere non-performative 'end' in itself, we suggest there must be a much deeper engagement with a range of the situated realities of tech workers. This requires turning away from the paradoxical 'narrow generalizations' (Amrute, 2019) that tend to manifest in the abstract principles of high-level AI ethics frameworks such as 'fairness and justice', and towards a set of relational practices that emphasise the development of shared lexicons and ethical strategies across domains. This also means that AI ethicists must grapple meaningfully with the ethical concerns that arise from practitioners' own narratives of their work, such as the inevitable partiality of algorithmic accounts, the centrality of DEI to ethical AI development and deployment, and the very real practical limitations that legacy systems and data detachment pose to the objectives of AI ethics. By drawing on feminist scholarship we have created a 'thicker' and more nuanced picture of tech workers' perspectives on the ethical issues that arise when developing and maintaining AI systems. This study thus not only reveals the importance of treating tech workers as ethical stakeholders in AI development and deployment, but also demonstrates the importance of feminist approaches to ethical AI.

Acknowledgements

Many thanks to Dr Youngcho Lee, our research assistant on the Gender and Technology project, for her invaluable assistance throughout the data collection process; Dr Stephen Cave, for his guidance on this project; Llinos Edwards, for copyediting; and the reviewers and editorial team at *Big Data & Society*, for their thoughtful and detailed feedback on our paper.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Christina Gaw.

ORCID iDs

Jude Browne https://orcid.org/0000-0003-2492-2627
Eleanor Drage https://orcid.org/0000-0003-1696-4536
Kerry McInerney https://orcid.org/0000-0002-0053-4657

Notes

1. We use the term 'tech worker' to mean those directly involved in the conception, design, marketing, use and sale of AI in a given context.
2. These themes, listed by the order in which appear most frequently, are: (1) *transparency*, or increasing the explainability and interpretability of AI systems; (2) *justice, equity and fairness*, which includes correcting algorithmic bias and the promotion of diversity and inclusion in the AI industry; (3) *non-maleficence*, or preventing AI from knowingly or predictably causing harm; (4) *responsibility or accountability*, which Jobin et al. (2019) note is 'rarely defined'; (5) *privacy*, again rarely defined, but is commonly connected to personal data security; (6) *beneficence*, or promoting human flourishing and well-being; (7) *freedom and autonomy*, which includes both positive freedoms such as self-determination and negative freedoms such as freedom from surveillance; (8) *trust*, which includes both the creation of a culture of trust among scientists and the trustworthiness of AI systems; (9) *sustainability*, or ensuring that AI is developed and deployed in environmentally sustainable ways; (10) *dignity*, which is never defined in frameworks and (11) *solidarity*, which emphasises the importance of a strong social safety net to protect against AI-powered products (Jobin et al., 2019: 391–396).
3. At this point the interviews had begun to offer repetitive responses, what Bryant and Charmaz (2007: 611) call 'the saturation point' whereby no new perspectives or codes are generated.
4. See note 4.
5. (#54 #35 #50 #39 #13 #40).

6. We shall return to this specific point on Wikipedia and others in our general discussion of evidence later in the article.
7. (#41 #62 #36).
8. (#24 #56 #59).
9. (#41 #24 #44 #8 #33 #18 #26 #60 #51 #48 #35 #6 #27 #43 #18 #30 #60 #42 #61 #45 #46).
10. (#48 #35 #6 #27 #43 #18 #30 #42 #61 #45).
11. (#19 #30).
12. (#59 #29 #46).
13. (#37 #59 #48 #34 #27 #50 #60 #23 #49 #26).
14. (#59 #48 #60 #23 #49)
15. (#13 #59 #62 #56 #34 #27 #39 #26 #36 #23 #61 #51 #25).
16. IBM, Google and Facebook, for example, have all developed initiatives to this effect: the AI Fairness 360 tool kit, the What-If Tool and Fairness Flow, respectively.

References

Ahmed S (2012) *On Being Included: Racism and Diversity in Institutional Life*. Durham, NC and London: Duke University Press.

Amoore L (2020) *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham, NC: Duke University Press.

Amrute S (2019) Of techno-ethics and techno-affects. *Feminist Review* 123(1): 56–73.

Baraitser L (2017) *Enduring Time*. London: Bloomsbury.

Bargu B (2013) Theorizing self-destructive violence. *International Journal of Middle East Studies* 45(4): 804–806.

Benjamin R (2019b) *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*. Durham, NC: Duke University Press.

Benjamin R (2019a) *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.

BMW (2020) Seven Principles for AI: BMW Group sets out code of ethics for the use of artificial intelligence. Available at: www.press.bmwgroup.com/global/article/detail/T0318411EN/seven-principles-for-ai:-bmw-group-sets-out-code-of-ethics-for-the-use-of-artificial-intelligence?language=en (accessed 16 November 2023).

Boyer A (2015) Data's Work is Never Done. *Guernica*, 13 March. Available at: www.guernicamag.com/anne-boyer-datas-work-is-never-done/ (accessed 8 March 2022).

Browne J (2023) AI & structural injustice: A feminist perspective. In: Browne J, Cave S, Drage E et al. (eds) *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*. Oxford: Oxford University Press, 328–346.

Browne J (2024) *Political Responsibility & Tech Governance*. Cambridge: Cambridge University Press.

Bryant A and Charmaz K (2007) *The SAGE Handbook of Grounded Theory*. London, UK: Sage.

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research* 81: 1–15.

Burt A (2020) Ethical Frameworks for AI Aren't Enough. *Harvard Business Review*, November 9. Available at: https://hbr.org/2020/11/ethical-frameworks-for-ai-arent-enough (accessed 21 November 2023).

Chan M, Lehmann J and Bundy A (2010) Higher-Order Representation and Reasoning for Automated Ontology Evolution. *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*. Available at: https://www.pure.ed.ac.uk/ws/portalfiles/portal/25316206/HIGHER_ORDER_REPRESENTATION_AND_REASONING_FOR_AUTOMATED_ONTOLOGY_EVOLUTION.pdf (accessed 16 November 2023).

Costanza-Chock S (2020) *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA: MIT Press.

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

Digital Catapult (2022) Ethics Framework. Available at: https://migarage.digicatapult.org.uk/ethics/ethics-framework/ (accessed 16 November 2023).

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

Drage E and Frabetti F (2023) The performativity of AI-powered event detection: How AI creates a racialized protest and why looking for bias is not a solution. *Science, Technology, & Human Values* 0(0): 1-28. DOI: 10.1177/01622439231164660.

Drage E and Mackereth K (2022) Does AI debias recruitment? Race, gender, and AI's 'eradication of difference'. *Philosophy and Technology* 35(4): 1–25.

European Parliament (Directorate-General for Parliamentary Research Services) (2019) A Governance Framework for Algorithmic Accountability and Transparency. Publications Office of the European Union. Available at: https://op.europa.eu/en/publication-detail/-/publication/8ed84cfe-8e62-11e9-9369-01aa75ed71a1/language-en (accessed 1 November 2022).

Fraser N (2016) Contradictions of capital and care. *New Left Review* 100: 99–117.

Gebru T (2019) Dealing with Bias in Artificial Intelligence. *The New York Times*, 2 January. Available at: www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html (accessed 14 March 2023).

Gray LM, Wong-Wylie G, Rempel GR, et al. (2020) Expanding qualitative research interviewing strategies: Zoom video communications. *The Qualitative Report* 25(5): 1292–1301. May.

Haas L and Gießler S (2020) In the realm of paper tigers – exploring the failings of AI ethics guidelines. *Algorithm Watch*, 28 April. Available at: https://algorithmwatch.org/en/ai-ethics-guidelines-inventory-upgrade-2020/ (accessed 21 November 2023).

Hagendorff T (2023) AI ethics and its pitfalls: Not living up to its own standards? *AI Ethics* 3: 329–336.

Hampton LM (2021) Black Feminist Musings on Algorithmic Oppression. Conference on Fairness, Accountability, and Transparency (FAccT '21), 3–10 March, Virtual Event, Canada. ACM, New York.

Hao K (2019) In 2020, let's stop AI ethics-washing and actually do something. *MIT Tech Review*, December 27. Available at: https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/. (accessed 21 November 2023).

Haraway D (1988) Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies* 14(3): 575–599. Autumn.

Haslanger S (2015) What is a (Social) Structural Explanation? *Philosophical Studies* 9 January. Available at: https://dspace.mit.edu/bitstream/handle/1721.1/97040/HaslangerWISSE-FINAL.pdf?sequence=1&isAllowed=y (accessed 1 November 2022).

Institute for the Future and Omidyar Network (2018) Ethical OS. Available at: https://omidyar.com/news/omidyar-network-partners-with-institute-for-the-future-to-launch-the-ethical-operating-system-a-guide-to-anticipating-the-future-impact-of-todays-technology/ (accessed 16 November 2022).

Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–399.

Jugov T and Ypi L (2019) Structural injustice, epistemic opacity, and the responsibilities of the oppressed. *Journal of Social Philosophy* 50(1): 7–27.

Kak A and West SM (2023) Confronting Tech Power. *AI Now Institute*. Available at: https://ainowinstitute.org/2023-landscape (accessed 16 November 2023).

Kerr A, Barry M and Kelleher JD (2020) Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society* 7(1): 1-12. DOI: 10.1177/2053951720915939.

Khan A, Akbar M, Fahmideh M, et al. (2023) AI Ethics: An Empirical Study on the Views of Practitioners and Lawmakers. IEEE Transactions on Computational Social Systems, 1–14. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10066257 (accessed 16 November 2023).

Lemieux ME, Zhang R and Tripodi F (2023) "Too soon" to count? How gender and race cloud notability considerations on Wikipedia. *Big Data & Society* 10(1): 1-14. DOI: 10.1177/20539517231165490.

Metzinger T (2019) Ethics washing made in Europe. *Der Tagesspiegel Online*, April 8. Available at: https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html (accessed 21 November 2023).

Mohanty CT (1989) On race and voice: Challenges for liberal education in the 1990s. *Cultural Critique* 14(Winter): 179–208.

Munn L (2022) The uselessness of AI ethics. *AI and Ethics* 3: 869–877.

NATO (2021) Summary of the NATO Artificial Intelligence Strategy. Available at: www.nato.int/cps/en/natohq/official_texts_187617.htm (accessed 1 November 2022).

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Nuttall L (n.d.) Five Technology Design Principles to Combat Domestic Abuse. IBM Corporation. Available at: www.ibm.com/policy/wp-content/uploads/2020/06/Design-Principles-to-Combat-Domestic-Abuse.pdf (accessed 8 November 2022).

OECD (2021) OECD AI's live repository of over 260 AI strategies & policies. OECD.AI, powered by EC/OECD. Available at: www.oecd.ai/dashboards (accessed 16 November 2023).

Osman K (2014) Paid Editors on Wikipedia – Should You Be Worried? *The Conversation*, 21 August. Available at: https://theconversation.com/paid-editors-on-wikipedia-should-you-be-worried-30527 (accessed 1 November 2022).

Pendergrast A and Pendergrast K (2021) A New AI Lexicon: MAINTENANCE. AI Now Institute, Medium. Available at: https://ainowinstitute.org/series/new-ai-lexicon (accessed 16 November 2022).

Peterson VS (2003) *A Critical Rewriting of Global Political Economy: Integrating Reproductive, Productive and Virtual Economies*. London: Routledge.

Phan T, Goldenfein J, Kuch D, et al. (2022) *Economies of Virtue: The Circulation of 'Ethics' in AI*. Amsterdam: Institute of Network Cultures.

Powell AB, Ustek-Spilda F, Lehuedé S, et al. (2022) Addressing ethical gaps in 'Technology for Good': Foregrounding care and capabilities. *Big Data & Society* 9(2): 1-12. DOI: 10.1177/20539517221113774.

Prescod-Weinstein C (2018) Diversity is a Dangerous Set-Up. Medium. Available at: https://medium.com/space-anthropology/diversity-is-a-dangerous-set-up-8cee942e7f22 (accessed 14 March 2023).

PWC (2019) A Practical Guide to Responsible AI. Available at https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf (accessed 16 November 2023).

Rességuier A and Rodrigues R (2020) AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society* 7(2): 1-5.

Rolls-Royce (2021) The Aletheia Framework Available at: www.rolls-royce.com/sustainability/ethics-and-compliance/the-aletheia-framework.aspx (accessed 1 November 2022).

Tripodi F (2021) Ms. categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society* 25(7): 1–21.

UNESCO (2021) General Conference, 41st. Report of the Social and Human Sciences Commission (SHS). Available at: https://unesdoc.unesco.org/ark:/48223/pf0000379920.page=14 (accessed 1 November 2022).

Ustek-Spilda F, Powell A and Nemorin S (2019) Engaging with ethics in Internet of things: Imaginaries in the social milieu of technology developers. *Big Data & Society* 6(2): 1-12. DOI: 10.1177/2053951719879468.

Vakkuri V, Kemell K-K and Abrahamsson P (2020) AI Ethics in Industry: A Research Framework. CEUR Workshop Proceedings. Available at: http://ceur-ws.org/Vol-2505/paper06.pdf (accessed 1 November 2022).

Walcott R (2019) The end of diversity. *Public Culture* 31(2): 393–408.

West SM, Whittaker M and Crawford K (2019) Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. Available at: https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2 (accessed 16 November 2023).

Whittaker M (2021) The steep cost of capture. *Interactions* 28(6): 50–55.

Wodeki B (2021) Deloitte: Women make up just 26 percent of AI Workforce in the US. AI Business, 5 October. Available at: https://aibusiness.com/document.asp?doc_id=769384 (accessed 1 November 2022).

Young E, Wajcman J and Sprejer L (2021) Where are the Women? Mapping the Gender Job Gap in A'. Policy Briefing: Full Report. The Alan Turing Institute. Available at: www.turing.ac.uk/research/publications/report-where-are-women-mapping-gender-job-gap-ai (accessed 1 November 2022).

Young IM (2000) *Inclusion and Democracy*. Oxford: Oxford University Press.

Zavalishina J and Polonski V (2017) Can We Teach Morality to Machines? Three Perspectives on Ethics for Artificial Intelligence. Oxford Internet Institute, 19 December. Available at: www.oii.ox.ac.uk/news-events/news/can-we-teach-morality-to-machines-three-perspectives-on-ethics-for-artificial-intelligence/#continue (accessed 16 November 2023).