Ethics and Information Technology (2024) 26:4
https://doi.org/10.1007/s10676-023-09739-1

ORIGINAL PAPER

Engineers on responsibility: feminist approaches to who's responsible for ethical AI

Eleanor Drage[1] · Kerry McInerney[1] · Jude Browne[2]

Published online: 2 January 2024

Abstract
Responsibility has become a central concept in AI ethics; however, little research has been conducted into practitioners' personal understandings of responsibility in the context of AI, including how responsibility should be defined and who is responsible when something goes wrong. In this article, we present findings from a 2020–2021 data set of interviews with AI practitioners and tech workers at a single multinational technology company and interpret them through the lens of feminist political thought. We reimagine responsibility in the context of AI development and deployment as the product of work cultures that enable tech workers to be responsive and answerable for their products over the long and short term. From our interviews, we identify three key pain points in understanding the distribution of responsibility between actors and developing responsible design and deployment practices: (1) unstable business ecosystems and AI lifecycles, which require an approach to responsibility that accounts for the dynamic nature of these systems; (2) the issue of incentivizing engineers to take responsibility for the mundane maintenance practices essential to the functioning of AI systems and (3) the need to overcome individual and structural barriers to taking ownership over AI products and their effects. From these findings, we make three recommendations based on feminist theory: (1) organisations should move from a static model of responsibility to a dynamic and ethically motivated *response*-ability; (2) companies need to revalue care and maintenance practices; and (3) firms must move away from individualistic ideas of responsibility towards fostering wider cultures of responsibility.

Keywords Artificial intelligence · Ethics · Feminism · Responsibility · AI ethics

Introduction

The wide scale development and deployment of artificial intelligence (AI) poses distinct challenges to organizational governance and oversight. In this paper, we define AI as algorithmic tools which are often, but not always, created through the use of machine learning (ML) techniques. The challenges AI poses range from explainability (the 'black box problem' in AI and ML) and difficulties in understanding how an AI-powered system has made a decision (Browne, 2023), through to issues relating to the unprecedented speed and scale at which AI technologies operate. Collectively, the question of who is responsible when AIgoes wrong has become increasingly urgent, and responsibility has been labeled one of the most important principles in AI ethics (Jobin et al., 2019). Responsibility for the erroneous outputs and harmful effects of AI systems becomes especially fraught when AI replicates, perpetuates, or exacerbates existing patterns of discrimination and injustice. Numerous high-profile instances of this – from Amazon's gender discriminatory AI-powered hiring tool to the UK government's A-level algorithm[1] – have increased public and political awareness of why AI requires responsible development, deployment and oversight. 'Responsible AI'

✉ Kerry McInerney
kam83@cam.ac.uk

[1] Leverhulme Centre for the Future of Intelligence, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, UK

[2] Centre for Gender Studies, University of Cambridge, Alison Richards Building 7 West Road, Cambridge CB3 9DT, UK

[1] In 2018, Amazon retired their AI-powered CV scanner after being unable to stop it from discriminating against female applicants (see Dastin, 2018). In 2020, the UK government used an algorithm to assist with allocating grades in A-level examinations. The government faced public outrage for using a tool that gave lower grades to state school students (Harkness, 2020).

has now become a brand-like umbrella term for the development of principles, approaches and methods of understanding what responsible AI development means and how it can be implemented.

States, international organizations and the technology sector have contributed 'Responsible AI' frameworks and initiatives to the AI ethics landscape, with Responsible AI both improving brand reputation (Burkhardt et al., 2019) and arguably becoming a brand in itself. The Council of Europe's study on responsibility in AI (2019) argues that society's concepts, institutions and practices of responsibility are critically important for ensuring that individuals and organizations are appropriately held to account for the adverse effects of AI. However, responsibility is rarely defined in the literature on Ethical AI, and the way in which the term is used varies depending on the context (Dastani & Yazdanpanah, 2022). As Christian Kästner argues, the exact responsibility that software engineers and data scientists have for their products is also contested (2022). There is an insufficient explanation of who is responsible and to whom, how responsibility is allocated and distributed, and what happens when responsibilities are not fulfilled.

We respond to this gap by bringing a feminist perspective to the question of what responsibility means and who should be involved. We address not only responsibility allocation, how it is distributed and what happens when it is not fulfilled, but also seek to reframe responsibility in feminist terms. This involves situating responsibility practices not as responses to a particular moment of crisis, but rather as constant modes of maintenance and care for AI systems and the wider social and political context in which these technologies are made and situated. We argue that responsibility must be vested in an acknowledgement of our mutual interdependence with human and nonhuman life (Braidotti, 2021). We use feminist theory to approach the concept of responsibility expansively. To do so, we draw on feminist theories of care and responsibility from Maria Puig de la Bellacasa, Donna Haraway, Iris Marion Young, and Judith Butler. Crucially, our feminist understanding of responsibility differs from leading paradigms of responsibility in organizational studies, which while foregrounding the need for 'responsiveness' in organizational governance, still largely center on a reactive, rather than proactive or pre-emptive, conceptualization of responsibility. This perspective aligns closely with organizational studies' emphasis on transforming governance structures to allow for greater agility or flexibility in technology development and deployment processes (Walter, 2021). Our feminist approach to responsibility in AI ethics is not premised on agility, but rather on a recognition of the inherent interdependence of human and planetary life, and our mutual and unchosen responsibility to safeguard and protect these relations.

We use feminist perspectives on responsibility to analyze a rich empirical dataset collected over the course of one year (2020–2021), where we interviewed tech workers about their views and approaches to AI and AI ethics. As Slota et al. (2021) note, the regulation of AI products often struggles to keep pace with innovation in the AI industry and the novel challenges posed by new technologies. This means that regulation by governmental and external bodies must be complemented by meaningful shifts within AI corporate culture to foreground conversations around responsibility and responsible AI development. This study offers empirical data about practitioners' observations on responsibility in relation to AI. It therefore speaks directly to practitioner concerns about responsibility and provides companies with increased capacity to identify problem areas and map meaningful steps forward.

The article is structured as follows: we open with an exploration of the existing literature on AI and responsibility, before laying out our specific feminist theoretical approach to AI. We then detail our methods and findings, and go on to discuss three major themes that emerged from tech workers' reflections on responsibility: (1) the barriers to responsibility posed by an ecosystem in flux; (2) the need to look at an AI product and its potential harms more holistically, rather than through the lens of 'tunnel vision', which is common in AI and data science; and (3) the structural impediments that prevent engineers from taking ownership over AI-powered products and their effects. We then explore the potential limitations of our feminist approach to responsibility, before concluding with what this study shows us about the state of AI ethics in corporate contexts, and how we can move forward from here.

Background: responsibility for AI

Our data concerns engineers' perceptions of responsibility attribution in AI and the creation of cultures of responsibility in companies that develop and deploy AI. We therefore exclude legal and regulatory literature on responsibility from this literature review.

Literature on responsibility beyond legislative work can be broadly divided into four categories: first, work that aims to encourage a safer and more responsible approach to AI development by attempting to model and ascribe responsibility in the development process (Dignum n.d.; Dastani & Yazdanpanah, 2022; Hedlund & Persson, 2022; Rochel and Evéquoz, 2021); second, philosophical work on the attribution of responsibility, largely moral philosophy approaches (Browne forthcoming, 2024; Kiener forthcoming; Bartneck et al., 2021; Tigard, 2021); third, surveys and studies that explore and determine responsibility allocation based on empirical findings (Association of Nordic Engineers 2021; IBM, 2022; Stack Overflow, 2018; Lu et al., 2021; Kelley, 2022); and fourth, explorations of the specific AI-related situations in which responsibility becomes a problem and how to communicate this to relevant stakeholders (IBM, 2022a; Taylor et al., 2018; Pratt, 2021; Brookings, 2021; Sebastián, 2021). Our study relates in particular to the second, third and fourth categories by offering feminist philosophical perspectives on responsibility, presenting findings from our study on how practitioners allocate responsibility and exploring perspectives on when and how responsibility becomes a problem for them. We aim to contribute different philosophical underpinnings to responsibility literature, which is currently dominated by moral philosophy approaches (e.g. Watson, 2004; Wallace, 1998, cited in Council of Europe, 2019). Established feminist work on responsibility offers an untapped resource centered around ethics that can reframe the debate.

Delineating work on responsibility from other branches of Ethical AI such as explainability or accountability is challenging, because while responsibility sits within the remit of Ethical AI, it is not Ethical AI's entire focus (UNESCO, 2020; Lu et al., 2021). The interchangeability of responsibility with adjacent terms such as accountability (IBM, 2022a) brings another challenge to scoping exercises that seek to define the parameters of responsibility literature. This exchanging of terms is made possible by the fact that responsibility is rarely defined (Jobin et al., 2019).

Theoretical framework

In this paper, we use feminist theory and methodological approaches to inform our approach to responsibility. Feminist theory has not only been applied to but has developed organizational studies' approaches to corporate responsibility, particularly by highlighting the relationship between corporate irresponsibility and masculinity in the context of, for example, the 2008 global financial crisis (Knights & Tullberg, 2012; for feminist approaches to the global financial crisis, see also True, 2012 and Hozic & True, 2016). While most feminist approaches to organizational studies largely focus on responsibility in relation to gender disparities in the workplace (Grosser & Moon, 2019; see also Browne, 2018, 2022), we build on work that uses feminist theory to inform and diagnose issues relating to corporate ir/responsibility that do not directly pertain to either gender or AI (Borgerson, 2007; Ferguson, 1997). Literature on corporate social responsibility provides a useful foundation given the importance of business ethics and corporate values to how AI is created, integrated into, and deployed by corporations (Browne et al., 2024). A recurring theme in feminist organizational studies is the desire for a responsive and care-minded definition of responsibility that extends beyond the common view that an employee is only responsible for what they directly create. Borgerson (2007), for example, emphasizes the importance of not only taking responsibility for what has been done or what should have been done, but for the ongoing conditional: what should be done.

This view intersects with selected work from ethics and technology scholars that signal the importance of responsiveness as a core component of responsibility (Van de Poel and Sand, 2018; Von Schomberg, 2012). Schomberg suggests that responsibility requires social actors and innovators that are "mutually responsive" and "interactive" (50), demonstrating the importance of flexibility and reactiveness when practicing responsibility as being sensitive to changes. This is particularly pertinent to technology development, where responsibility for a product is the result of many actions and actors, and is therefore a distributed phenomenon (Coombs et al., 2003; Rothwell, 1994). Organizational and strategic technology studies literature have responded to these diagnoses by advocating for the implementation of, for example, a bottom up approach to empirical investigation; an acknowledgement of how users shape and transform innovations; participatory design; and the combating of organizational silos (Harty, 2010).

Feminist scholarship has long grappled with what it means to live ethically within myriad kinds of socio-technical systems, from institutions and languages to human bodies - themselves an assemblage of complex, technological and societal forces (McNeil, 2007). We base our conceptualization of responsibility on two key principles or ideas that emerge from feminist theory: 1) the interdependence of human life and the ethical responsibilities that arise from this (Puig de la Bellacasa, 2012; Butler, 2003); and 2) the need to understand human actions and undertakings in relation to broader social and political structures (Chambers, 2008), particularly structural injustice (Young, 2011; Browne, 2023). In their 2003 essay 'Violence, Mourning, Politics', feminist philosopher Judith Butler argues against the idea that gender is a discrete and autonomous category that individuals own, foregrounding instead how gender is fundamentally a relational category: "neither gender nor sexuality is precisely a possession, but, rather, is a mode of being dispossessed, a way of being for another or by virtue of another" (2003, p. 13). More broadly, they ask whether ontological interdependence could provide the ethical basis for a politics based on and oriented towards common human vulnerability: "Although I am insisting on referring to a common human vulnerability, one that emerges with life itself, I also insist that we cannot recover the source of this vulnerability: it precedes the formation of 'I'" (Butler, 2003, p. 20). Nonetheless, Butler also insists that this vulnerability is differentially distributed among peoples, groups and countries. While Iris Marion Young also observed an unequal distribution of vulnerability, she nevertheless claimed that all of us share a responsibility for the harms that emanate from the inevitable interconnectedness and structural relations of common human environments (albeit to differing degrees) (Young, 2011). We frame our feminist understanding of responsibility through this emphasis on both (involuntary) interdependence and the unequal distribution of vulnerability.

The emphasis on interconnectedness and situatedness similarly shapes feminist approaches to science and technology studies (STS). From Donna Haraway's thinking in relation to the game of cat's cradle and the situatedness of knowledge production, and Karen Barad's theory of agential realism to Susan Leigh Star's onion thought experiment and Judy Wajcman's analysis of manufactured expectations of the future, feminist STS's close examination of sociotechnical systems has been central to the establishment of feminist AI ethics (Haraway, 1988, 1994; Barad, 2007; Star, 1990; Adam, 1998; Wajcman, 2010). Catherine D'Ignazio and Lauren Klein draw on the sum of this work to argue that technologies are not discrete or independent variables, but rather reflective of the wider systems and structures in which they are developed and deployed (D'Ignazio & Klein, 2020). Meanwhile, philosophers and political theorists such as Browne (2023, 2024) and Kasirzadeh (2022) lay out innovative new approaches to tech governance based on Iris Marion Young's theories of structural injustice and responsibility. Strategic feminist interventions in AI ethics embrace a pro-justice politics to strengthen technology and organizational scholarship by arguing that participatory design is often not meaningfully inclusive, and de-biasing strategies should not attempt to make systems neutral but rather foreground how and where products and actors are situated in relation to one another and how this results in harmful effects (Kalluri, 2020; Drage & Mackereth, 2022). For example, Drage and Frabetti (2023) have argued that AI performatively produces the world it claims to observe according to the interests of its stakeholders. Acknowledging AI's worldmaking capabilities means focusing on how systems are situated in relation to corporate and shareholder interests and market factors, and therefore how bias is better expressed as the partial perspectives taken by AI systems, which are shaped by systems of power (Browne et al., 2024).

A better understanding of the contingent and context-specific expression of harm requires greater insight into how employees who work with AI draw on their own expertise and professional experience to grapple with everyday challenges posed by the attribution of responsibility. This, we believe, is a necessary corrective to overarching AI ethics frameworks and responsible AI programmes that primarily rely on high-level abstractions and universalising ethical principles, rather than being embedded within and emerging from specific contexts (Browne et al., 2024). Furthermore, while the majority of our interviewees are engineers and data scientists, we also interviewed employees from a wide range of sectors within the organization, as we detail in the methods section of this paper. Slota et al. (2021) have noted that AI engineers should not be the exclusive knowledge-keepers of AI, and it should be understood that a range of stakeholders across and beyond an organization possess important insight into how a particular system functions and also what its effects might be when deployed. Knowledge about AI is embedded and produced through a complex web of relations between institutional culture, employees, users and vendors. Therefore, in this study we interviewed a sample of workers for whom AI is a part of their daily work in order to expose and explore what they consider to be their personal responsibility in designing AI systems.

Methods

This study is based on a set of 63 qualitative interviews conducted at a large multinational tech company located in the Global North. To retain the anonymity of our research partner, we refer to them here as TCX. We have taken other measures to protect the company's privacy and that of its employees, omitting any names, geographical locations or references to specific tech systems. Between 2020 and 2021 we interviewed 63 employees online from a number of teams including Marketing, Legal, AI Research, AI Design and Human Resources. We interviewed as broad a cross-section of the organization as possible, from graduate to senior employees and across a range of demographic indicators (age, self-defined gender, ethnicity and registered disabilities). A point person at TCX helped us establish the research partnership and contact our first interviewees; the rest of our interviewees were selected using the snowballing method. Hence, while we were able to interview a demographically diverse dataset from different parts of the organization, our interviewee selection may have been influenced by those already predisposed to a project on gender, feminism and technology.

While we asked our interviewees at TCX about a wide range of issues relating to AI ethics, this paper is concerned with responses to the following question posed to these practitioners: who is responsible for AI? We purposefully did not define responsibility for interviewees in order to better ascertain what responsibility meant to them and avoid pre-empting their responses. As we elaborate elsewhere, we explicitly drew on feminist research methods in designing and conducting our interviews (Browne et al., 2024). For example, we went to lengths to ensure our interviewees' anonymity and used semi-structured methods to allow them to tell their own 'stories' about AI ethics. Our dataset for this paper consists of 56 references to responsibility in 27 separate interviewee files. NVivo was used for coding and to assist with data analysis. We used cross-coder analysis to verify our coded data, which demonstrated four main themes: ecosystems in flux, ownership, tunnel vision and responsible engineers. We subsequently offer four recommendations, relating to response-ability, owning up, maintenance and framing responsibility beyond crisis.

Findings from Our study: who is responsible, how, and why?

Practitioners are divided about who should be responsible for particular AI systems, how this could be implemented and why responsibility practices were necessary. Some employees were unsure where responsibility did and should lie (#59 #9 #38 #27 #25). Of those that expressed views and suggestions, respondents explained that responsibility for AI was either shared across the organization, split between multiple employees with distinct ownership of particular elements of a system's operation, or located in one particular role or person. Some suggested it should be a combination, citing the multiple actors or teams involved in product creation and deployment as a reason why no single actor could take full responsibility (#41 #35 #47 #25 #49 #59 #26 #40). For example, Interviewee 40 insisted that:

> It's a shared responsibility amongst the producers of the data […] [t]he model builder, the people using the model, the people providing the tools to build the models, the people creating the tools to host them and [those who] maintain the models in operation. (#40)

The 'shared' category often felt that responsibility was a moral question and that 'across the piste, everyone needs to take more responsibility' (#49), or that engineers should themselves be moral people (#35). Some gave conflicting responses that expressed a desire to both foster a shared sense of responsibility and establish precise ownership (#56 #49). Others suggested that responsibility was not so much shared as split (#47 #62 #34 #51 #63) between different teams or employees who had responsibility for different kinds of systems: '[T]here is a difference between the teams that support the infrastructure and those that support the logic of what the bot does. And I think it's understanding that split of responsibility.' (#62).

Another 'split' in responsibility was identified between AI engineers that represent and are represented by the company, and external auditors. This point of view was best represented by Interviewee #34, who said:

> The burden of proof on showing that you've done your due diligence should lie with the AI engineers and software developers, the company effectively, but the validation of that […] should lie with an external body […] similar to an OFCOM for AI.

Respondents looked to existing corporate metaphors of the 'chain of management' or 'chain of custody' to explore a connected, orderly and distributed model for attributing *different kinds* of responsibility. For Interviewee #51, the 'chain of custody' connected the different places where responsibility was 'located':

> At every point we need to be capturing who is responsible and who's doing what. And at any moment in the project, if you're a data scientist you should be able to look upstream and identify who the stakeholder was in the chain of custody one link up from you. (#51)

This linear view of responsibility resonated with engineers as not only an ordered and systematic approach that aligns with their own disciplinary backgrounds but also the suggestion of a community of responsible stakeholders in constant contact with each other. This sense of reciprocity and connectedness, signified by the chain, was extremely important to practitioners. Others argued that the 'layered models' of AI should be reflected in a layered approach to responsibility, where 'you understand […] who is responsible for the functions within [each layer of the model]' (#62). However, the downsides to distributed responsibility were also expressed by respondents, notably that a lack of individual ownership was resulting in a 'random' distribution of 'spread' responsibility (#36). This was perceived to be the cause for a shirking of responsibility and 'panic' when crises occur: 'At the moment the responsibility is not clear […] something goes wrong, people panic normally, and they start pointing fingers' (#38). Respondents claimed that responsibility only became a focal point when things were going wrong, and that this was leading to an inability to assign responsibility. They suggested that one way of avoiding the inevitable dodging of responsibility that arose from only considering responsibility when faced with problematic situations was to install clear and collective organizational responsibility. This might take the form of, for example, 'responsible AI eyes' that provide oversight during product development stages (#51), or an ethics committee' (#36) that could flag when engineering teams come up with ideas that ‘contain bias’ (#51) and rectify them to avoid AI-generated harms. As we will discuss later, respondents were concerned that locating responsibility with a committee rather than engineering teams might constitute a siloing or outsourcing of responsibility.

This is likely why many respondents felt that responsibility is something that can or should be ‘located’, and that there was a place where ‘the buck stops’ – whether in one place or many others (#41 #34 #36 #39 #27 #44 #2 #25 #63 #23 #61 #49 #38 #51 #33). That place or role included the people that deploy AI systems, engineers (the responsibility would be with the people that made the machine learning model, right?’ #39), product owners (‘I'm quite clear that the ultimate responsibility lies with the product owner’ #61), and ‘the business end’ (#63 #23 #47), which included the CTO and CEO (‘I guess the buck stops ultimately with […] the CEO’ #63). Several respondents even stated that the purpose of senior employees was to take final ownership of AI systems: ‘That's where the rise of the CTO has come [from] as well, you've tried to [create] that figurehead, that board level person, who effectively would own it, and someone there would go “That's where the buck stops”’ (#23).

Those who placed the burden of responsibility on the people deploying systems or deciding how they should be used also claimed that the way in which data and AI systems are made actionable has the greatest impact in making AI harmful or beneficial. They indicated that the interpretation of results generated by particular kinds of systems was just as crucial as the system itself: ‘I think [responsibility] is largely about how [AI is] used, […] how whatever data the AI system is providing, or insight, is actually acted upon […] So ultimately it comes down to that person or that group of people, making those decisions’ (#41). Understandably, it was largely AI engineers that took this view, not necessarily because they wanted to deflect responsibility from themselves but because they were aware that their work did not take place in a vacuum. This provided insight not only into immediate technical responsibility practices but also how the wider corporate infrastructure could demonstrate more responsibility for its technologies and support engineers in this objective. This will be evidenced in the subsections below.

Discussion

We have identified three main themes in relation to how the wider corporate infrastructure should support responsible AI and how engineers can be enabled to practice responsibility as they build AI systems. These are: (1) unstable business ecosystems and AI lifecycles, which require an approach to responsibility that accounts for the dynamic nature of these systems; (2) the issue of incentivizing engineers to take responsibility for the mundane maintenance practices essential to the functioning of AI systems; and (3) the need to overcome individual and structural barriers to taking ownership over AI products and their effects.

An ecosystem in flux: moving from scattered responsibility to response-ability

A key issue that emerged in many conversations was that companies are not stable but always in flux. Interviewees emphasized how high employee turnover and staff moving between teams posed a significant barrier to identifying who was responsible for a particular system (#41, #2, #21). Respondents claimed that the creation of products by people who left the company soon after resulted in a subsequent lack of ownership over those products: ‘You just find things that are orphaned, they have no owner, because somebody left and, and there wasn't a process for tying a new person in and making sure there was someone responsible’ (#51). ‘Orphaned’ products were seen as an unhappy consequence of employee turnover rather than an inevitable part of the product life cycle. Engineers felt that acting responsibly should mean embedding explainability mechanisms into AI which communicated a model's risks to employees deploying systems in the future: “You need ways to make sure that you can send messages downstream to people who might use your models in the future, so that they are warned about your view of how this model could be used safely” (#51). When this did not happen and information about how a model worked was not kept up to date, it was seen to break the aforementioned “chain of custody”: “the chain of custody is broken for all these artifacts because we haven't got really tight mechanisms for keeping those things up to date” (#51). Interviewees stated that a broken chain of custody created separations between systems and their original creators. These separations could be temporal, spatial, or disciplinary: “something that has been built 30 years ago is now being implemented in an organization on the other side of the world” (#59). This is a common phenomenon; the accelerated modularization of technology development has resulted in “the disaggregation of development responsibility” and “the diffusion of responsibility for ethical outcomes” (Wessel and Helmer, 2020, pp. 51). Interviewees identified these degrees of separation as leading to uncertainty about who was responsible: “So is it the implementer of the AI who is responsible, the person who has put it into their business process? To be honest with you I don't know” (#59).

The above concern about the lack of a responsible individual or entity was often also connected to product ownership (#48 #27 #47 #23 #51 #25 #49 #20 #33 #62 #44 #38 #36 #9). Respondents who mentioned ownership saw it as "key" to the problem of responsibility (#23). Through the lens of ownership, interviewees viewed responsibility as lying both with the business and with engineers:

> There are different levels of responsibility in AI. You have the responsibility for the system itself […], [engineers] have the responsibility to build tools and build solutions like that. But there is a business ownership of the data, and there is a business ownership of the outcome as well. (#47)

This testimony aligns with respondents who identified multiple owners at different stages (#51 #23). However, these multiple stages were also the stated reason for the lack of ownership, as we have previously explored in relation to orphaned systems and products shared across groups. Another more surprising issue suggested as a reason why employees lacked ownership for their systems was the box-ticking ethics frameworks which encouraged them to clear themselves of responsibility, and therefore of ownership:

> People have to take ownership and in the frameworks you just almost tick them through a checkoff […] I've ticked the box and therefore I'm clear [of] responsibility. If I ask someone in my team to do something, I want them to own the outcome and not own the task. (#20)

Respondents had a sense that responsibility meant more than one-off bias checks, and that this short-termism was preventing practitioners from "owning" their tools. They connected responsibility to more than just adhering to protocol ("the task") but with an enduring commitment to technologies and their downstream effects ("the outcome'". For Interviewee #20, "owning the outcome" means considering the impact of employees' tools rather than just their immediate functionality. But being able to consider these outcomes in advance requires, in the words of other interviewees, "stepping back", and educating themselves on the effects of different kinds of systems.

The flow of people in and out of a corporation is paralleled by the movement of systems and products across porous organizational boundaries and was similarly identified as a challenge to the attribution of responsibility. The movement of products within large corporate systems, as well as the layering of systems on top of each other, makes it difficult to identify who is responsible for products or datasets (Browne et al., 2024). According to some AI engineers, once a system has been developed it belongs to the people who "look after the models", prompting a shift in responsibility. One respondent suggested that "Lots of people I know [who build AI] don't want to be responsible for that because it really should be the responsibility of the teams that look after these models" (#2), and that "Sometimes [responsibility] can be fed backwards and then you're stuck doing debugging, which you don't really want to be doing, you want to be doing something more interesting" (#2). This metaphor of forwards- or backwards-moving responsibility was commonly used by engineers who felt that the AI lifecycle makes it impossible to attribute permanent responsibility to one team. AI engineers were keen to have this mobility reflected in the attribution of responsibility, to avoid not only personal culpability but also, as this quote evidenced, the extra – and unprestigious – labor of fixing a problem that had been passed "back" down the line to them.

Many engineers exhibited a desire to practice and promote more responsibility for the products they were creating. In light of this, feminist STS can be used to inform a different conception of responsibility, one tailored to the problems posed by distributed ecosystems in flux. For example, feminist STS work on response-ability can be used to reconfigure the relationship between responsibility and liability. The goal here is to enable engineers to practice responsibility without fear of individual culpability by engaging in a 'cultivating collective knowing and doing' or 'making-with' each other [sympoiesis] (Haraway, 2016, pp. 34, 58). This intervention takes place at the level of business culture, which should encourage and reward employees for being attuned to their environment, customers and contexts. Chicana activist and feminist theorist Gloria Anzaldúa argues that "the ability to respond is always what is meant by responsibility" (1999, p. 42). As we explore in the next section, this is also contingent on feeling comfortable and supported enough to acknowledge harm, seek redress for this harm, and communicate issues to senior leadership. Anzaldúa directs us towards an ethos of responsibility where employees are incentivised and equipped with the necessary tools with which to build and deploy AI responsibly.

This framing can help address the problems that arise when static and individualistic models of responsibility fail to map cleanly against the in-flux ecosystems identified by our interviewees. Instead of seeking out culpable individuals, which treats responsibility as primarily an individual property, engineers should collectively enable each other to practice responsible AI development. This means thinking about responsibility as relational: dependent on positive and enabling relationships which direct attention towards the potentially affected persons. As Coeckelbergh (2020) argues, too much attention on the agents practicing responsibility often deflects from what should be the actual focal point – those that suffer when crises occur. Ensuring that the communities and individuals who are most likely to be harmed are placed at the center of corporate discussions about responsibility requires stepping back from a singular focus on the technical operations of a product and looking more holistically at the wider social and political context within which this product is developed and deployed. We explore the need to 'step back' in the following section. While this approach is limited by its reliance on companies being invested in responsible business culture, as demonstrated by our study, many tech companies are motivated to go further than meeting legal obligations; more advanced understandings of responsibility in AI ethics can assist them in their endeavor.

Seeing the bigger picture: changing incentives, revaluing care work

In order to counter the tunnel vision that accompanies an intense focus on the way a system functions in laboratory conditions at a particular point in time, some interviewees pointed to the importance of 'stepping back' to see the bigger picture (#49 #51 #55). They suggested that AI engineering facilitated tunnel vision when it came to assessing the impact its systems would have down the line. This meant that engineers were unable to consider factors beyond technical performance:

> Today, a lot of the emphasis on model monitoring is really just about performance metrics, […] the assumption is it's working fine. There'll be a bit of drift in the environment, and then I'll need to retrain it. And that strikes me as a very limited view of what might go wrong with the model. (#51)

In other words, technical fixes were seen as limited when it came to exercising responsibility for, in particular, adverse effects that systems might have on customers. While these testimonies showed that some engineers were concerned by being unable to access 'the bigger picture', they also revealed that there were currently no mechanisms in place for them to do this on the job. In fact, they explained that the fast-paced environment of tech development was not conducive to this approach. They cited agile and other 'speedy' ways of working as a reason why it was difficult to take time to think about the bigger picture.

> Engineers are often scapegoated because they're doing the work. And if you think about it, when you've got someone standing over you with a mallet, not literally, […] you just focus on doing [the task at hand] and you're not able to think anymore [about the bigger picture] because agile's a super speedy process. (#55)

Pressurized work environments clearly affected engineers' capacity to consider elements of tech development beyond the immediate task. In other words, their ability to think responsibly was being constrained by a focus on resolving technical problems. Ultimately, the implication was that these kinds of irresponsible work environments were leading to irresponsible engineers.

Our interviewees highlighted how they were not asked to think more holistically about the potential harms of their products, nor were they asked to consider how they could take responsibility for these products and their effects over time. This, we argue, is a result of maintenance practices being insufficiently valued. AI maintenance requires constant recalibration, management and tests to avoid algorithmic drift and ensure systems are still working as intended. As the practitioners in this study noted, it also requires knowing when to 'switch off' systems that are no longer needed or working as required (Browne et al., 2024). Being responsible for an AI system means not only being held accountable when it goes wrong, but engaging in routinized care and maintenance. Hence, as Kelly Pendergast and Anna Pendergast write, "to bring an AI system into the world is also to bring about a responsibility for its care" (2021). Yet, feminist theorists and economists have extensively demonstrated how caring responsibilities are distinctly gendered and racialized forms of work; they are differentially distributed along gendered and racial lines, ill-respected, and poorly remunerated (Peterson, 2003; Fraser, 2016; Puig de la Bellacasa, 2012; Atanasoski & Vora, 2019).

Inspired by scholarship in feminist political economy, we suggest that engineers will only be incentivized and able to practice greater responsibility over AI systems when maintenance work is revalued as necessary and important work. The entire economy of AI goods and systems is highly dependent on forms of undervalued maintenance work. Critical scholarship in the field of AI ethics has foregrounded the forms of arduous and repetitive labor that are frequently outsourced to gendered and racialized workers, to the Global South and even (in the case of Finland) to prisoners, further demonstrating the ever-expanding reach of the prison-industrial complex (Chen, 2019). These tasks include data labeling, the cleaning and maintenance of data sets, and content moderation (Posada, 2021; Gray & Suri, 2019; Bui, 2022; D'Ignazio & Klein, 2020). Similarly, Erin Young, Judy Wajcman and Laila Sprejer's (2021) study on gender inequality in the AI industry shows that women are more likely to fulfill roles and tasks relating to dataset cleaning and maintenance rather than the more prestigious roles in machine learning, leading them to argue that the

AI industry is characterized by persistent structural gender inequality.

We propose to push this critical feminist work forward in the field of AI ethics by centering the ordinary maintenance of AI systems in our understanding of where responsibility should lie for AI systems. We recommend that companies provide greater recognition at both an individual and structural level for maintenance work. Rather than punishing individual scapegoats when a system goes awry, companies need to revalue the work of maintenance in order to prevent such situations arising in the first place. This may require moving from a paradigm of individual ownership to one of collective stewardship and responsibility for the care of AI systems.

Revaluing maintenance labor also requires rethinking the temporal lifetime of AI systems. This means not only considering when systems should be switched off, as mentioned earlier, but also grappling more intentionally with the impact of AI systems post-development. The impacts of technological systems on people and the environment shift over time, and these shifts can be brought to light through the lens of maintenance (Pendergast and Pendergast, 2021). While the Silicon Valley-esque hype around AI and the constant pursuit of profit drives a culture of excessive innovation, we argue that a feminist approach to responsibility requires slowing down and looking after old systems as well as making new ones. Rather than fetishising creation, novelty and newness, AI engineers must focus on how to practice responsibility for AI systems in ordinary and ongoing timeframes.

Currently, debates around responsibility in relation to AI systems tend to emerge in moments of acute crisis. However, as feminist theorists of temporality demonstrate, this focus on the moment of crisis ignores the lapses of care and failures of responsibility that routinely occur in the temporal registers of the ordinary: failures that are "ordinary, chronic, and cruddy rather than catastrophic, crisis-laden, and sublime" (Povinelli, 2011, p. 3). In the process of AI development and deployment, this might look like the failure to consider alternative points of views; harms that may arise from technologies in the initial stages of development process; the failure to attend to how gender, race and other systems of power shape data collection and usage, even when these datasets appear to be neutral or free from gendered and racialized markers (see Browne et al., 2024); a lack of consideration for minoritised customers' needs and desires in product development; the assumption that users can be represented by a universalized user persona (Costanza-Chock, 2020); and, as discussed in the previous section, the failure to ensure that systems continue to operate as intended and that their accuracy does not decrease.

While we may only see the effects of these failures of responsibility and care when they emerge at the moment of crisis – such as the Amazon hiring algorithm or the A-level algorithm – they are fundamentally underpinned by a persistent undercurrent of irresponsibility. Thus, being responsible for AI systems requires, in the words of Lisa Baraitser (2017), 'taking care of time': taking care of systems in the ordinary times, not just the extraordinary ones. This approach may also help combat the 'tunnel vision' identified above and assist engineers in 'stepping back' to see the wider impact of their products.

Barriers to ownership: owning it, owning up

Several respondents suggested that responsible tech begins with responsible engineers. But as with the above, others were concerned that business practices were not enabling engineers to practice responsibility. An example of this was the view that a system's secondary uses did not always align with engineers' views about where their systems should or shouldn't be deployed (#37 #51 #17):

> I'm building a platform that could be used for things that I really don't agree with, and that's something I have to think about […] So who's responsible if something's used [in a way] the creator didn't intend it for? Who's to blame? Someone has to be to blame so it's definitely […] a big concern. (#17)

Further to the evident anxiety about colleagues and managers pointing fingers when systems were misused, Interviewee #17 also suggests that if the tool is deployed for an unintended purpose then they are no longer responsible for it. Therefore, ongoing responsibility is contingent on using an AI system in a way that is consistent with its intended use at the time of development. Engineers flagged that people making business decisions about particular AI systems might not have the requisite knowledge to decide how systems can be repurposed; they would need to trust engineers to outline what particular systems should be used for and then adhering to these recommendations (#47 #55). Decisions about AI therefore need to be collaborative and informed, using engineer expertise. These respondents' testimonies suggest that an ethics committee might not be sufficient in ensuring that systems are made and deployed ethically because they would not have first-hand knowledge of the systems they were overseeing. While some still indicated that responsibility for a product should also lie with the business (or high-level decision-makers), they also believed that these stakeholders might not understand how the product works or has been built, and this would limit the efficacy of their responsibility practices. They also suggested that this would additionally inhibit engineers from being able to work responsibly, because decisions were being taken out of their hands by a group of people that were inadequately equipped to make them.

Another respondent that pointed to barriers in engineers being able to exercise responsibility raised the issue of hiring ethical technologists (#56). They suggested that responsible AI development requires ethical people: 'in part, the engineer [is responsible] ethically for making the tool, so it's then [about] the person's ethics and morality and how they operate that' (#56). For them, responsible engineering required a moral workforce. This view is largely absent from the responsibility literature, which focuses on moral systems rather than moral engineers.

In response, we suggest that AI firms should work towards fostering a wider culture of responsibility among engineers and tech workers. For example, postcolonial philosopher Gayatri Spivak explains that being responsible is about being responsive and answerable (2003: 102). It means 'owning up' and creating a proximity to and knowledge of affected parties rather than 'looking around' in an effort to delegate responsibility. Spivak also highlights the reparative dimensions of responsibility, looking to the climate crisis to show that our responsibility for the planet is sustained through active and concerted sustainability efforts. Just as it is the planet's right to expect collective responsibility, it is the consumer's right to expect companies to practice responsible AI creation that extends beyond the attribution of culpability. Reparative responsibility in the climate context means actively reversing climate change, and in AI deployment this means creating products and work processes that actively benefit marginalized groups, as has been attempted through, for example, algorithmic reparations (Davis et al., 2021). Responsibility here means undertaking the urgent reparative action required to move towards social justice as we build technologies (Spivak, 2003, p. 48).

However, it is crucial to note that owning up is not merely about individual employees acting responsibly. Instead, organizations are responsible for building work cultures where employees feel comfortable enough to own up without being afraid of being scapegoated. As Sarah Franklin notes, ethical technologies beget ethical research cultures, where researchers and engineers experience psychological safety, and feel able to ask difficult ethical questions and consider alternative different points of view (Franklin, 2022). This, she argues, is a basic premise of feminist scholarship and methods – asking who is heard in a room, and whose voice goes unheard (Franklin, 2022). Thus, responsibility means creating research and work cultures that are attuned to how systems of power can make being responsible far more difficult and dangerous for some employees than others. It also requires thinking about responsibility as a collective rather than an individual endeavor.

Conclusion and limitations

With this paper, we aim to demonstrate the utility and insights of feminist theorising in relation to engineers' own positions on responsibility and what it means to be responsible for an AI-powered product within a complex corporate ecosystem. Our rich empirical dataset of interviews with practitioners clearly demonstrates the need for further conversations at industry level about what responsibility is in order to meaningfully implement AI ethics protocols and agendas. Our findings suggest that practitioners currently struggle with how responsibility is attributed and practiced within industry settings. We have shown that feminist theory offers a way of rethinking responsibility as a relational principle that requires business contexts to support answerable and responsive actors. These feminist approaches allow practitioners to grapple with responsibility not only in relation to risk and harms but also to ongoing maintenance practices. Crucially, this work is not merely additive (suggesting additional approaches to responsibility) but invites a re-centering of multiple conceptions of responsibility. These feminist approaches, we argue, can facilitate the creation of responsible AI development and deployment cultures.

We recognise the limitations and meaningful challenges that engineers and other tech workers may face when attempting to put these feminist approaches to responsibility into practice. AI systems are rarely 'located' within single enterprises, crossing private, for-profit companies and different types of entities and organizations with dramatically different motivations, priorities, and frameworks of understanding (Keyes & Austin, 2022). Therefore, while it is important to explore the individual cultures of responsibility at single technology firms, this research must be complemented by further exploration of what responsibility for AI could and should look like across sectors and organizations. Moreover, when drawing on the lens of critical feminist political economy, we must also question the extent to which engineers' capacity to be ethically responsive to the short- and long-term consequences of their products is limited by the for-profit context in which they operate. As D'Ignazio suggests, within corporate structures the overarching model of accountability is "accountability to shareholders and profit"; and that creating more responsible AI cultures would require "reining in that profit motive that is extractive, exploitative, and exacerbating systemic inequality" (2021). There is a significant risk that engineers and tech workers will simply lack the necessary incentives to enact and practice feminist-informed conceptions of responsibility within the present for-profit corporate structure. Consequently, cultural shifts within tech firms must also be accompanied by the effective enforcement of existing regulatory tools to ensure that companies consider the potential harms of their products and successfully mitigate them (Kak & West, 2023; Browne, 2024). Additionally, without broader pro-justice cultural transformations within technology firms, it is possible that the labour of being responsible for or caring for AI products and their effects will disproportionately fall on junior workers or structurally disadvantaged staff. Hence, while we aim to shed light on the utility of feminist thinking on responsibility for transforming AI development and deployment cultures, we also recognise that responsibility remains a crucial and contentious issue for the field of AI ethics. We thus recommend further investigation into how AI firms can develop ethical cultures of responsibility informed by feminist work and scholarship, while simultaneously advocating for broader structural transformations of the AI industry towards justice.

Acknowledgements Many thanks to Dr Stephen Cave, Director of the Institute of Technology and Humanity, and Dr Youngcho Lee, our research assistant on the Gender and Technology project. Many thanks also to the editorial team and reviewers at Ethics and Information Technology for their kind and thoughtful feedback. Finally, we would like to thank our interviewees for their generosity and their time.

Declarations

Conflict of interest We have no conflicts of interest to disclose.

References

Adam, A. (1998). *Artificial Knowing: Gender and the thinking machine*. Routledge.

Anzaldúa, G. (1999). *La frontera/Borderlands*. Aunt Lute Press.

Association of Nordic Engineers (2021). *Addressing Ethical Dilemmas in AI: Listening to Engineers*. Available at: https://nordicengineers.org/wp-content/uploads/2021/01/addressing-ethical-dilemmas-in-ai-listening-to-the-engineers.pdf.

Atanasoski, N., & Vora, K. (2019). *Surrogate humanity: Race, Robots and the politics of Technological futures*. Duke University Press.

Barad, K. (2007). *Meeting the Universe halfway: Quantum Physics and the entanglement of Matter and meaning by Karen Barad*. Duke University Press.

Baraitser, L. (2017). *Enduring Time*. Bloomsbury.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). *Ethics in AI and Robotics: A Strategic Challenge*. Springer. https://doi.org/10.1007/978-3-030-51110-4_12. Cham.

Borgerson, J. (2007). On the harmony of Feminist ethics and business ethics. *Business and Society Review*, *112*(4), 477–509.

Braidotti, R. (2021). Posthuman Feminism and Gender Methodology in Jude Browne (Ed.) (2021) Why Gender? Cambridge: Cambridge University Press. pp 101–125.

Brookings (2021). What exactly is 'responsible AI' in principle and in practice? Conference, 10 May. Available at: www.brookings.edu/events/what-is-responsible-ai/.

Browne, J. (2018). Technology, Fertility and Public Policy: A structural perspective on human egg freezing and gender Equality. *Social Politics*, *25*(2), 149–168. https://doi.org/10.1093/sp/jxx022.

Browne, J. (2022). The corporate Father. *Globalized fatherhood* (Vol. 27, pp. 31–52). Berghahn Books. https://doi.org/10.1515/9781782384380-004.

Browne, J. (2023). AI and Structural Injustice: A Feminist Perspective. In *Feminist AI*:.

Browne, J. (2024). *Political responsibility & tech governance*. Cambridge University Press. Forthcoming.

Browne, J. (Ed.). Stephen Cave, Eleanor Drage and Kerry McInerney. Oxford: Oxford University Press. pp 328–346.

Browne, J., Drage, E., & McInerney, K. (2024). forthcoming Understanding Corporate Tech Workers' Perspectives on Ethical Issues in AI Development: Foregrounding Feminist Approaches. Big Data & Society.

Bui, L. T. (2022). *Model machines: A history of the Asian as Automaton*. Temple University Press.

Burkhardt, R., Hohn, N., & Wigley, C. (2019). Leading your organization to responsible AI. Quantum Black, 2 May. Available at: www.mckinsey.com/capabilities/quantumblack/our-insights/leading-your-organization-to-responsible-ai.

Butler, J. (2003). Violence, mourning, Politics. *Studies in Gender and Sexuality*, *4*(1), 9–37. https://doi.org/10.1080/15240650409349213.

Chambers, C. (2008). *Sex, culture and justice: The limits of choice*. Penn State University Press.

Chen, A. (2019). Inmates in Finland are training AI as part of prison labor. The Verge. Available at: www.theverge.com/2019/3/28/18285572/prison-labor-finland-artificial-intelligence-data-tagging-vainu.

Coeckelbergh, M. (2020). Artificial Intelligence, responsibility attribution, and a relational justification of Explainability. *Science and Engineering Ethics*, *26*, 2051–2068. https://doi.org/10.1007/s11948-019-00146-8.

Coombs, R., Harvey, M., & Tether, B. (2003). Distributed processes of provision and innovation. *Industrial and Corporate Change*, *12*, 1051–1081.

Costanza-Chock, S. (2020). *Design Justice: Community-led practices to build the worlds we need*. MIT Press.

Council of Europe (2019) A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Prepared by (MSI-AUT). Available at: https://rm.coe.int/responsability-and-ai-en/168097d9c5.

D'Ignazio, C. (2021). Catherine D'Ignazio on Data Feminism. The Good Robot. Available at https://www.thegoodrobot.co.uk/post/data-feminism-with-catherine-d-ignazio (accessed 4/12/2023).

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press.

Dastani, M., & Yazdanpanah, V. (2022). Responsibility of AI systems. *AI & Soc*. https://doi.org/10.1007/s00146-022-01481-4.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, 11 October. Available at: www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Davis, J. L., Williams, A., & Yang, M. W. (2021). Algorithmic reparation. *Big Data & Society*, 8(2). https://doi-org.ezp.lib.cam.ac.uk/10.1177/20539517211044808.

de la Puig, M. (2012). Nothing comes without its World': Thinking with Care. *The Sociological Review*, *60*(2), 197–216. https://doi.org/10.1111/j.1467-954X.2012.02070.x.

Dignum, V. (n.d.) Responsible Artificial Intelligence (Ed.). Available at: https://ec.europa.eu/jrc/communities/sites/default/files/03_dignum_v.pdf.

Drage, E., & 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). https://doi.org/10.1177/01622439231164660. https://journals.sagepub.com/doi/.

Drage, E., & Mackereth, F. (2022). Does AI Debias Recruitment? Race, gender, and AI's 'Eradication of difference'. *Philosophy & Technology*, *35*(4), 89. https://doi.org/10.1007/s13347-022-00543-1.

Ferguson, K. E. (1997). Postmodernism, Feminism, and organizational ethics: Letting difference be. *Women's studies and business ethics. Towards a new conversation. Ed. A L Larson and R E Freeman*. Oxford University Press.

Franklin, S. (2022). Sarah Franklin on Reproductive Technologies and Feminist Research Ethics. *The Good Robot*. Available at https://www.thegoodrobot.co.uk/post/sarah-franklin-on-reproductive-technologies-and-feminist-research-ethics.

Fraser, N. (2016). Contradictions of Capital and Care. *New Left Review*, *100*, 99–117.

Gray, M., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from Building a New Global underclass*. Harper Business.

Grosser, K., & Moon, J. (2019). CSR and Feminist Organization Studies: Towards an Integrated theorization for the analysis of gender issues. *Journal of Business Ethics*, *155*, 321–342. https://doi.org/10.1007/s10551-017-3510-x.

Haraway, D. J. (1988). Situated knowledges: The Science question in Feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575–599.

Haraway, D. J. (1994). A game of Cat's cradle: Science studies, Feminist Theory, Cultural studies. *Configurations*, *2*(1), 59–71.

Haraway, D. J. (2016). *Staying with the trouble: Making kin in the Chthulucene*. Duke University Press.

Harkness, T. (2020). A Level Results: Why Algorithms Aren't Making the Grade. Available at: www.sciencefocus.com/future-technology/a-level-results-why-algorithms-arent-making-the-grade/amp/.

Harty, C. (2010). Implementing innovation: Designers, users and actor-networks. *Technology Analysis & Strategic Management*, *22*(3), 297–315. https://doi.org/10.1080/09537321003647339.

Hedlund, M., & Persson, E. (2022). Expert responsibility in AI development. *AI & Soc*. https://doi.org/10.1007/s00146-022-01498-9.

Hozic, A., & True, J. (Eds.). (2016). *Scandalous economics: Gender and the politics of financial crises*. Oxford University Press.

IBM (2022). Responsibility for AI Ethics Shifts from Tech Silo to Broader Executive Champions, says IBM Study. Available at: https://newsroom.ibm.com/2022-04-14-Responsibility-for-AI-Ethics-Shifts-from-Tech-Silo-to-Broader-Executive-Champions,-says-IBM-Study.

IBM (2022a). Accountability. Available at: www.ibm.com/design/ai/ethics/accountability/.

Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics guidelines. *Nat Mach Intell*, *1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Kak, A., & West, S. M. (2023). Confronting Tech Power. AI Now Institute. Available at https://ainowinstitute.org/2023-landscape (accessed 22/11/2023).

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature. Available at https://www.nature.com/articles/d41586-020-02003-2?utm_ source=twt_nnc&utm_medium=social&utm_campaign=naturenews&error=cookies_not_supported&code=bcb71409-8168-4ee3-98f5-8e7578698393.

Kasirzadeh, A. (2022). Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. P. 349–356. https://doi.org/10.1145/3514094.3534188.

Kästner, C. (2022). Responsible AI Engineering. Medium, 7 Jan. Available at: https://ckaestne.medium.com/responsible-ai-engineering-c97e44e6c57a.

Kelley, S. (2022). Employee perceptions of the effective adoption of AI principles. *Journal of Business Ethics*, *178*, 871–893. https://doi.org/10.1007/s10551-022-05051-y.

Keyes, O., & Austin, J. (2022). Feeling fixes: Mess and emotion in algorithmic audits. *Big Data and Society*, *9*(2). https://doi.org/10.1177/20539517221113772.

Kiener, M. (forthcoming) (Ed.). Can We Bridge AI's Responsibility Gap at Will? *Ethical Theory and Moral Practice* 25 (4):575–593.

Knights, D., & Tullberg, M. (2012). Managing masculinity/mismanaging the corporation. *Organization*, *19*(4), 385–404. https://doi.org/10.1177/1350508411408170.

Lu, Q., Zhu, L., Xu, X., Whittle, J., Douglas, D., & Sanderson, C. (2021). Software Engineering for Responsible Ai: An Empirical Study and Operationalised Patterns. Data61, CSIRO, Sydney, 19 November. https://arxiv.org/abs/2111.09478.

McNeil, M. (Ed.). (2007). *Feminist Cultural studies of Science and Technology, London*. Roudedge.

Pendergrast, A., & Pendergrast, K. (2021). A New AI Lexicon: Maintenance. AI Now, Medium, 22 June. Available at: https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-maintenance-6a653bccdb99.

Peterson, V. S. (2003). *A critical rewriting of global political economy: Integrating Reproductive, productive and virtual economies*. Routledge.

Posada, J. (2021). A New AI Lexicon: Labor. AI Now, Medium, 23 September. Available at: https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-labor-2db30b6b2349.

Povinelli, E. (2011). *Economies of Abandonment: Social belonging and endurance in late Liberalism*. Duke University Press.

Pratt, M. (2021). AI Accountability: Who's Responsible when AI Goes Wrong? Tech Target, 19 August. Available at: www.techtarget.com/searchenterpriseai/feature/AI-accountability-Whos-responsible-when-AI-goes-wrong.

Rochel, J., & Evéquoz, F. (2021). Getting into the engine room: A blueprint to investigate the shadowy steps of AI Ethics. *AI & Soc*, *36*, 609–622. https://doi.org/10.1007/s00146-020-01069-w.

Rothwell, R. (1994). Towards the fifth-generation innovation process. *International Marketing Review*, *11*(1), 7–31.

Sebastián, M. (2021). First-person representations and responsible Agency in AI. *Synthese*, *199*, 7061–7079. https://doi.org/10.1007/s11229-021-03105-8.

Slota, S. C., Fleischmann, K. R., Greenberg, S., et al. (2021). Many hands make many fingers to point: Challenges in creating Accountable AI. *AI & Soc*. https://doi.org/10.1007/s00146-021-01302-0.

Spivak, G. (2003). *Death of a Discipline*. Columbia University Press.

Stack Overflow (2018). Developer Survey Results. Available at: https://insights.stackoverflow.com/survey/2018/.

Star, S. L. (1990). Power, Technology and the Phenomenology of conventions: On being allergic to onions. *The Sociological Review*, *38*(1), 26–56. https://doi.org/10.1111/j.1467-954X.1990.tb03347.x.

Taylor, S., Pickering, B., Boniface, M. (2018). Responsible AI – Key Themes, Concerns & Recommendations for European Research and Innovation. HUB4NGI, 2 July. Available at: https://zenodo.org/record/1303253#.Y4UOpHbP1D8.

Tigard, D. W. (2021). Responsible AI and Moral responsibility: A common appreciation. *AI Ethics*, *1*, 113–117. https://doi.org/10.1007/s43681-020-00009-0.

True, J. (2012). *The Political Economy of Violence Against women*. Oxford University Press.

UNESCO (2020). Virtual Discussion of the Ad Hoc Expert Group (AHEG) for the Preparation of a Draft Text of a Recommendation on the Ethics of Artificial Intelligence. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000373199.

van de Poel, I., & Sand, M. (2018). Varieties of responsibility: Two problems of responsible innovation. *Synthese*, *198*(19), 4769–4787.

Von Schomberg, R. (2012). Prospects for Technology Assessment in a framework of responsible research and innovation. In M. Dusseldorp, & R. Beecroft (Eds.), *Technikfolgen abschätzen Lehren: Bildungspotenziale transdisziplinärer Methoden* (pp. 39–61). Springer.

Wajcman, J. (2010). Further reflections on the sociology of technology and time: A response to Hassan. *The British Journal of Sociology*, *61*, 375–381. https://doi.org/10.1111/j.1468-4446.2010.01317.x.

Wallace, R. J. (1998). *Responsibility and the Moral sentiments*. Harvard University Press.

Walter, A. T. (2021). Organizational agility: Ill-defined and somewhat confusing? A systematic literature review and conceptualization. *Manag Rev*, *71*, 343–391. https://doi.org/10.1007/s11301-020-00186-6.

Watson, G. (2004). Two Faces of Responsibility. *Agency and Answerability: Selected Essays*. Oxford, 2004; https://doi.org/10.1093/acprof:oso/9780199272273.003.0010.

Wessel, M., & Helmer, N. (2020). A Crisis of Ethics in Technology Innovation. *Sloan Review*, https://sloanreview.mit.edu/article/a-crisis-of-ethics-in-technology-innovation/.

Young, I. M. (2011) *Responsibility for justice*. Oxford: Oxford University Press.

Young, E., Wajcman, J., & Sprejer, L. (2021). *Where are the women? Mapping the gender job gap in AI*. Full Report. The Alan Turing Institute. Available at: www.turing.ac.uk/research/publications/report-where-are-women-mapping-gender-job-gap-ai.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.