A feminist data ethics of care framework for machine learning: The what, why, who and how
by Joanne E. Gray and Alice Witt

Abstract
This article conceptualises and provides an initial roadmap for operationalising a feminist data ethics of care framework for the subfield of artificial intelligence (‘AI’) known as ‘machine learning’. After outlining the principles and praxis that comprise our framework, and then using it to evaluate the current state of mainstream AI ethics content, we argue that this literature tends to be overly abstract and founded on a heteropatriarchal world view. We contend that because most AI ethics content fails to equitably and explicitly assign responsibility to actors in the machine learning economy, there is a risk of implicitly reinforcing the status quo of gender power relations and other substantive inequalities, which contribute to the significant gap between AI ethics principles and applied AI ethics more broadly. We argue that our feminist data ethics of care framework can help to fill this gap by paying particular attention to both the ‘who’ and the ‘how’, as well as by outlining a range of methods, approaches, and best practices that societal actors can use now to make interventions into the machine learning economy. Critically, feminist data ethics of care is unlikely to be achieved in this context unless all stakeholders, including women, men, and non-binary and transgender people, take responsibility for this much needed work.

Contents

1. Introduction
2. What and why: A feminist data ethics of care framework for machine learning
3. Who: Humans in the machine learning economy
4. How: Operationalising feminist ethics of care for machine learning
5. Conclusion and future research

1. Introduction

Calls for feminist interventions into artificial intelligence (‘AI’) have grown louder and more numerous in recent years (Bassett, et al., 2020; D'Ignazio and Klein, 2020; Neff, 2020). Many of these calls seek to move beyond the AI ethics conversation (Crawford, et al., 2019; Metcalf, et al., 2019), by which we mean discussions about the ethical implications of “systems that display intelligent behaviour by analysing their environment and taking actions with some degree of autonomy to achieve specific goals” (European
A feminist data ethics of care framework for machine learning: The what, why, who and how

Commission, 2021). These interventions are much needed given the arguably limited impact of AI ethics content in practice (Crawford, et al., 2019) and the relative dearth of practical guidance around applied AI ethics (Morley, et al., 2020), especially for those seeking to investigate and address specific problems like racism, sexism, or classism.

Feminist scholars have made significant inroads in developing diverse theories, methods and best practices for data studies; however, their focus to date has largely been on social media, as opposed to AI. In this article we seek to extend this work, drawing particularly on the ground-breaking efforts of Luka and Leurs (2020), who in turn draw on the work of their colleagues (see e.g., Bivens, 2015; Luka and Millette, 2018), to explore what feminist data ethics for AI could look like in practice. Specifically, we aim to develop a feminist data ethics of care approach for the subfield of AI known as ‘machine learning’ (‘ML’). We define machine learning as algorithms, or models, that analyse and draw inferences from patterns in large datasets for human and/or automated decision-making, involving varying degrees of human instruction or supervision. The limited scholarship on feminist approaches to machine learning is significant given the “explosion” of this technology in everyday life, from recommender systems (e.g., videos on YouTube and Netflix), to personalised products (e.g., Apple smartphones and watches), targeted advertising (e.g., when scrolling through content on Instagram) and law enforcement (e.g., police use of facial recognition software).

While this technology can produce a range of benefits, it can also cause serious harm when it is used in a way that, for example, unfairly discriminates against people based on gender, class or race. For instance, when recidivism risk software suggest people of colour are highly likely to commit future crimes, regardless of individual criminal records (Angwin, et al., 2016). Harms can arise from issues at any stage in the machine learning pipeline — the often step-by-step processes of data collection and preparation, system development, and deployment of machine learning enabled technologies (Suresh and Guttag, 2021). Common issues include biases embedded during data collection and classification (Alvi, et al., 2018), but they can also occur due to choices made during the design, development and deployment of machine learning infrastructure and technologies (Hasselbalch, 2021). The machine learning economy — the infrastructure, processes, organisations and individuals involved in the development and deployment of machine learning technologies — is therefore an important site of study.

The aim of this article is to conceptualise and provide an initial roadmap for operationalising a feminist data ethics of care framework for the machine learning economy. In Section 2, we outline our feminist data ethics of care framework that is underpinned by two main principles: machine learning, like other types of AI, is “partial, situated and contextual” (Haraway, 1988; Leurs, 2017); and a range of representational and allocative harms can arise in the machine learning pipeline (Barocas and Selbst, 2016). These principles are the foundation of our feminist data ethics of care praxis for preventing harm that focuses on: (i) ensuring diverse representation and participation in the machine learning economy; (ii) critically evaluating positionality; (iii) centring human subjects at every stage of the machine learning pipeline; (iv) implementing transparency and accountability measures; and (v) equitably distributing the responsibility for operationalising a feminist data ethics of care praxis. The latter is particularly important given that limited attention has been paid in AI ethics scholarship to the significance of gender power relations and the often-gendered nature of ‘care work’ (Organisation for Economic Co-operation and Development [OECD], 2018).

In Section 3, we evaluate the state of AI ethics content against our feminist data ethics of care framework, arguing that it tends to be overly abstract and is often founded on a heteropatriarchal world view. We contend that because most of these initiatives fail to explicitly assign responsibility for operationalising AI ethics, they risk implicitly reinforcing the status quo of gender power relations and other substantive inequalities, which contribute to the significant gap between AI ethics principles and applied AI ethics more broadly. In Section 4, we argue that our feminist data ethics of care framework can help to fill this gap by paying particular attention to both the ‘who’ and the ‘how’, with a view to achieving the equitable distribution of responsibility for operationalising feminist data ethics of care in the machine learning economy. Our framework also outlines key methods, approaches, and best practices that can be used now to
make feminist data ethics of care interventions into the machine learning economy.

Overall, we argue that feminist data ethics of care is unlikely to be achieved in practice unless all stakeholders, including women, men, and non-binary and transgender people, take responsibility for this work. While our framework focuses on machine learning, we contend that it has purchase across other subfields of AI, including neural networks, robotics and expert systems (Bullinaria, 2005). We conclude, in Section 5, by setting an ongoing research agenda for feminist data ethics of care in the machine learning economy and beyond.

2. What and why: A feminist data ethics of care framework for machine learning

There is no universal feminist theory or practice. As Cifor, et al. (2019) explain, “feminism is plural; there are many feminisms and they may differ in their positive visions, methodologies, collective ends, and situated concerns” [1]. Feminists are thus “an enormously diverse group of people with varying opinions” shaped by Black, queer, trans, Latinx, Indigenous and other lived experiences [2]. While feminist discourses are contested, many are united in action against gender subordination (Henry, et al., 2021) and in “a refusal of inheritance” [3] of the heteropatriarchal status quo across contexts and social domains. In line with these ideals, we conceptualise feminism in terms of achieving the substantive equality of rights, opportunities, responsibilities and outcomes for women [4], men, nonbinary, and transgender people. This is a pressing policy area given that “gender inequality persists everywhere and stagnates social progress” (United Nations, n.d.), to the detriment of women, girls and nonnormative identities especially (see, e.g., United Nations, 2019).

More specifically, our feminist perspective is founded on ethics of care, which many scholars recognise as an inherently feminist approach to ethics (D’Olimpio, 2019; Luka and Millette, 2018). In contrast to dominant moral theories that emphasise “rationality”, “logic” and a state of being unemotional, often in line with purportedly objective rules (see e.g., Bentham, 1996; Kant, 1993; Mill, 1859), ethics of care theorists argue for a contextualised and relational approach (Raghuram, 2019). Rejecting the dichotomy between reason and emotion, such theorists often apply ethical rules and undertake the work of caring for others in a manner involving emotions, including compassion and empathy. Within an ethics of care framework, “caring” is “a species of activity that includes everything we do to maintain, contain and repair our ‘world’ so that we can live in it as well as possible” [5]. This includes humans managing multilayered and intersecting relationships with themselves, others, states, institutions, the physical world and technologies.

According to Luka and Millette (2018), contemporary data ethics of care emerged in the 1980s from “second- and third wave feminist psychology, sociology, and cultural studies”, emphasising the importance of integrating “feminist and intersectional values into considerations of data analyses” [6]. Intersectionality is a “heuristic term” [7] that focuses attention on social, identity and ideological forces that affect and legitimize power (Crenshaw, 2017). An intersectional approach challenges single-axis thinking; for example, one that solely focuses on gender without considering how multiple structural forces such as race, gender and class can intersect for differently situated individuals (Crenshaw, 1990). This is important because social categories are “always permeated by other categories, fluid and changing, always in the process of creating and being created by dynamics of power” [8]. In an age of ever-increasing datafication, data is an important source of and vehicle for the exercise of power (Hasselbalch, 2021), which leads us to propose a data ethics of care approach for machine learning that integrates feminist and intersectional values. As illustrated in Figure 1, our framework comprises two overarching principles and a five-pronged praxis that we argue should guide attempts to develop, maintain and repair (Elish and Watkins, 2020) machine learning technology.
Our first principle is that machine learning, like other types of AI, is “partial, situated and contextual” [9]. It is now widely recognised that the data from which machines learn and produce information are not neutral (O’Neil, 2016). While machine learning technology is meant to think like “us”, it is in fact predicated on the world views, experiences and assumptions of its creators who are situated, subjective and often discretionary actors (Broad, 2018). Within the machine learning pipeline, “human expertise intervenes between raw data and the analysis, crucially shaping the data, the choice of analysis, and in some cases the truth claims associated with the analysis” (Muller, et al., 2019). Critically, practitioners involved in developing machine learning technologies make “data decisions” that can have negative downstream effects in high stakes domains, often with “outsized effects on vulnerable communities and contexts” [10]. Take, for example, skin cancer detection models that work less effectively for people with darker skin tones (Kinyanjui, et al., 2020; see generally Buolamwini, 2017).

When analysing machine learning systems, we advance Haraway’s (1988) “view from a body, always a complex, contradictory, structuring, and structured body, versus the view from above, from nowhere, from simplicity” [11]. This view is particularly important in the context of AI because it challenges the myth of disembodied objectivity, something the technology industry has long promoted in conjunction with notions of detached, impartial, mechanical and smart automated systems (Broad, 2018). This is not to reinforce an objective-subjective binary but rather to underline the fact that knowledge produced by machine learning systems is positional — derived from real and often limited human experiences — and subject to discretionary decision-making. The machine learning pipeline relies upon inputs of data created by humans.
— increasingly involving labour from low paid workforces in developing nations — and, to varying degrees, this data can embody the views, biases and ideologies of its creators (Royer, 2021). An example is when individuals or teams assign binary gender labels to people who may not identify as such (Access Now, 2021). Humans also exercise agency when designing and deploying machine learning enabled systems. For these reasons, machine learning technology is always situated and contextual, rather than objective and impartial.

This leads to our second feminist data ethics of care principle that machine learning can produce a range of representational and allocative harms at all stages of the pipeline. As Barocas and Selbst (2016) explain, *representational harm* refers to choices that consciously or unconsciously diminish particular identities through biased representations in data. *Historical bias* occurs when data reflects and fails to correct for structural inequities, such as when a dataset that captures low female representation in leadership is used to reinforce an unfounded heteropatriarchal belief that men are better leaders than women (Roselli, *et al.*, 2019). *Representation bias* can occur when training datasets under or over represent certain groups of individuals: for example, racist and sexist representations of black women and girls in search engines (Noble, 2018). Measurement bias can occur when machine learning models select or measure features in a dataset to determine outcomes that are discriminatory (Mehrabi, *et al.*, 2021). For instance, when a model uses race as a proxy for recidivism in law enforcement without accounting for social factors, such as the over-policing of certain communities.

Representational harms can lead to allocative harms and distributive injustices, including an over or under distribution of resources and opportunities to a particular group based on irrelevant personal characteristics such as race, gender, and sexual orientation (Lloyd, 2018). For example, some facial analysis or body scanning systems have been trained using “pale male” and binary-gender datasets that can lead to airport security systems singling out, or working less optimally for, bodies that do not align with white, heteronormative standards [12]. These biases are important issues of concern as machine learning models are increasingly used to support decision-making in high-stakes contexts, including education, banking, employment and housing (AI Now Institute, New York University, 2018).

It is not enough, however, for us to simply identify and discuss these harms as “principles alone cannot guarantee ethical AI” [13]. AI ethics work must form the foundation of affirmative action. We therefore advance a feminist data ethics of care praxis, based on the foregoing principles, aimed at helping actors in the machine learning economy to recognise the situated nature of machine learning and, critically, to empower them to act to prevent harm. By “praxis”, we mean ways to enact feminist data ethics of care in practice, as part of a “continuous loop of ethical research, social action/practice, and critical in-the-moment reflection or reflexivity as well as post-action evaluation” [14]. The five components of our praxis all work towards harm prevention: (i) ensuring diverse representation and participation in the machine learning economy; (ii) critically evaluating positionality; (iii) centring human subjects at every stage of the machine learning pipeline; (iv) implementing transparency and accountability measures; and (v) equitably distributing the responsibility for operationalising a feminist data ethics of care praxis. We discuss each of these in turn.

The first component of our praxis is concerned with the extent to which diverse populations are involved in the machine learning economy and empowered to act in the interests of those represented. A major cause for concern is that STEM (science, technology, engineering, mathematics) disciplines lack diversity (Women in AI Ethics™, 2020) and, more specifically, the development of AI systems are led by a “homogenous tribe” [15] of mostly affluent, highly educated, white American or Asian men (Chang, 2019). Some technology companies have attempted to promote diversity and inclusion through workplace training, employment quotas and other measures (Chang, 2019). However, as Webb (2020) points out, “talking about diversity — asking for forgiveness and promising to do better — isn’t the same thing as addressing diversity within the databases, algorithms, and frameworks that make up the AI ecosystem” [16]. The ‘leaky pipeline’ — a metaphor often used to describe the disappearance of women from STEM degrees and careers — and Silicon Valley’s “bro culture” further entrench the problem (Chang, 2019).
A feminist data ethics of care framework for machine learning: The what, why, who and how

A related concern is that many Western technology companies promote a “postfeminist sensibility” [17] that can make feminism seem like “second nature and therefore also unnecessary” [18]. A key driver of this sensibility is the dominant individualistic ideology of the West, including discourses of individual freedom, choice, empowerment and responsibility, which can give the false impression that the work of feminism is done. The reality is that we have not yet achieved gender parity: the outlook for “closing the global gender gap has increased by a generation from 99.5 years to 135.6 years” since the start of the COVID-19 pandemic (World Economic Forum, 2021). Of the 156 countries included in the Global Gender Gap Index, “women represent only 26.1% of some 35,500 parliament seats and just 22.6% of over 3,400 ministers worldwide”, which underlines the significantly limited extent that women exercise governing power (World Economic Forum, 2021). In Silicon Valley, in particular, representation of women on boards and in executive positions remains low (Bell and Belt, 2019). The danger of leaving the design and development of technologies that structure society to an elite, homogenous group (Firestone, 1970) is that they are unlikely to be fully attuned to the conditions of racism, sexism, classism and other intersecting oppressions. Addressing intersectional oppression arguably requires a shift from neoliberal individualism to a feminist consciousness and solidarity that challenges the institutional and cultural structures of power that affect the lives of everyday people.

The second component of our praxis involves actors in the machine learning economy critically examining their positionality and that of others. We might ask, for instance, who is collecting, cleaning, sorting, processing and reviewing data along the machine learning pipeline and what biases could these decision-makers introduce or reproduce. These are important lines of inquiry because, as we explained in the context of our first overarching principle, data are never “natural” or “raw” renderings of reality (see, e.g., Gitelman, 2013). Understanding positionality in a feminist context is also important given the “political whiteness” [19] of mainstream or “popular” (Banet-Weiser, et al., 2020) feminisms that can result in societal actors co-opting movements to the detriment of minority groups (see, e.g, the discussion of #MeToo in Phipps, 2020). Women of colour, among others, continue to call for non-Western and intersectional perspectives that examine how gender inequality intersects with other categories of social difference or identity, such as race, ethnicity, class, nationality, immigration, sexuality, age and disability, to produce structural and systemic forms of inequality (see e.g., Cho, et al., 2013; Collins, 2019; Crenshaw, 2017). By adopting intersectional approaches to positionality, we can map the machine learning economy in ways that do not gloss over cultural, social and other differences.

The third component of our feminist data ethics of care praxis involves centring human subjects at every stage of the machine learning pipeline. The rhetoric of automation and big data is seductive: it promises, for instance, efficient, instantaneous and infinite knowledge and to tackle, or even to solve, the world’s most pressing challenges (Mayer-Schönberger and Cukier, 2013). Not only is this technologically deterministic rhetoric optimistic at best, it can also promote the idea of disembodied data subjects, which decentres humans and their diverse concerns, lived experiences and power relations (Leurs, 2017). We argue for a human-centric approach that involves care for humans, particularly marginalised groups and individuals who are often disproportionately affected by machine learning gone wrong (Neff, 2020), and non-humans, such as animals and the natural environment (Merchant, 1980). This approach is important given the socio-technical nature of machine learning: that is, society shapes the design, development and deployment of technology, just as technology shapes society, including norms about the place of different people in the world (Elish and Watkins, 2020). Proponents of a feminist data ethics of care must be aware of the mutually constitutive relationship between technology and society and the potential for both to bring about change. In practice, actors in the machine learning economy can engage with community stakeholders in the data collection, development and deployment of machine learning technologies, bringing into focus the human and natural context and impact of their work at each stage of the machine learning pipeline. There should also be an emphasis on the individuals, teams and executives who make decisions throughout the machine learning pipeline — to avoid an overly objective or disembodied view of machine learning as a purely technical phenomenon.

Transparency and accountability measures are also integral to our praxis. Transparency, at its core, is “concerned with the quality of being clear, obvious and understandable without doubt or ambiguity”
(Belgium v Commission, 2005). Transparency should be prioritised throughout the pipeline, from the aims of a model to data provenance, analysis and deployment, and the documentation of day-to-day decision-making by coders and developers, including select coding languages. Developers might seek to mitigate harm through algorithmic impact assessments or in-depth risk-benefit assessments, interdisciplinary collaborations with researchers and other subject matter experts, and community stakeholder engagement. To move beyond the mere provision of information, it is vital that machine learning actors also implement accountability measures, including standards against which conduct and outcomes can be evaluated, to hold those who exercise power to account (Suzor, et al., 2019). These measures, among others outlined in Section 4, can help inform and improve societal understandings of the data and systems from which humans and machines generate meaning.

The fifth component of our feminist data ethics of care praxis is the equitable responsibility for and distribution of work. This starts with consideration of how, historically, paid work and care work (“work/care regimes”) have been organised along gendered lines (Raghuram, 2019). While the gendering of work is “situational and dynamic” [20], historical power relations continue to influence contemporary work/care regimes in which “caring” is feminised and women often undertake a disproportionate amount of care and unpaid domestic work, as distinct from men’s largely paid work (Fisher and Tronto, 1990; Jordan, 2020). Even when women undertake paid work, they often face a “‘double shift’ — the combination of a paid job and unpaid domestic work” [21]. This inequality is well-documented: for example, the United Nations found that in 2019 women spent “three times as many hours as men each day in unpaid care and domestic work” [22]. Our purpose in highlighting the often-gendered ordering of work/care regimes is to warn of the risk of women, and non-binary and transgender people, along with those who identify as BIPOC [23], undertaking a disproportionate amount of feminist data ethics of care work. As such, we argue that key to the widespread adoption of our proposed approach throughout the machine learning economy is the explicit and equitable assignment of responsibility to all stakeholders. Equitable distribution in this context requires that responsibilities for achieving feminist data ethics of care ends is integrated into day-to-day work practices, across teams and from the least to most senior employees, rather than siloed in standalone teams or projects. Before exploring how this might be achieved in practice, in Section 4, we first evaluate the emerging conventions of mainstream AI ethics content against our feminist data ethics of care framework.

3. Who: Humans in the machine learning economy

Despite an abundance of AI ethics content, there is little evidence to suggest that it has had a tangible impact on machine learning in practice (Hagendorff, 2020). Several convergent factors might account for this apparent lack of efficacy. First, the field of AI ethics is highly fragmented, with AI ethics content taking various forms including principles, frameworks, charters, manifestos, statements and guidelines (Jobin, et al., 2019; Mittelstadt, 2019). AI ethics content is also produced by a range of actors including government agencies, private companies, academic institutions and civil society organisations (Mittelstadt, 2019). These actors sometimes disagree about ethical theories, as well as whose interests should be served by ethical principles, how and to what ends (Mittelstadt, 2019). This means there is wide scope for stakeholders to prioritise different normative assumptions, interests and practical requirements across public and private contexts. Critically, AI ethics content is almost always voluntary and without legally binding obligations (Jørgensen, 2017), which leaves it open to varying levels of adoption. Thus, while AI ethics content might be plentiful, it lacks precision, remains open to debate and does little to compel participants in the machine learning economy to act. Indeed, Wagner (2018) warns of “ethics washing”: the practice of making non-binding commitments to ethics as a way to be shielded from public regulation, while eliminating the need for substantive action. A lack of transparency within the machine learning economy also makes it difficult to evaluate whether commitments to AI ethics have extended from theory to practice.

The apparent lack of efficacy of AI ethics content might also be explained by the types of principles that
tend to be included in these documents. In Jobin, et al.’s (2019) content analysis of 84 AI ethics documents, the principles transparency, justice and fairness, nonmaleficence, and responsibility were the most common. This set of principles appears to align fairly well with our feminist data ethics of care praxis at face value. Yet, on closer examination, it falls short both conceptually and pragmatically. For example, the principle of transparency, as articulated in AI ethics content, tends to be defined in terms of interpretability of data and “explainability” of AI enabled decision-making (Jobin, et al., 2019). There are calls for increased disclosure of algorithms and data, access to source code and capacity for third party audits. But the question of who transparency is for is largely unaddressed. If transparency is conceived in terms of access to data and technical systems, there is a risk that the utility of transparency will largely accrue to a heteropatriarchal class of technically proficient data scientists. Meaningful transparency must extend beyond technical access and should involve the disclosure of decision-making and outcomes throughout the machine learning pipeline for evaluation by a diverse range of actors.

A feminist data ethics of care approach to transparency might involve systematically documenting and analysing the positionality of the people involved in the design, develop and deployment of machine learning technology. The question of who is making decisions matters for understanding the potential biases and limitations that decision-makers embed within machine learning enabled systems. In addition to improving capacity for predicting and preventing harm, by evaluating the positionality of individual decision-makers, machine learning designers and deployers may also have an opportunity to improve the diversity of people making decisions throughout the pipeline. Within a feminist data ethics of care framework, transparency is not simply a technical issue, but one that aims to: identify opportunities for improving workplace diversity; identify real decision-makers in the machine learning pipeline to understand the potential biases they are at risk of embedding; and hold real actors to account (Hasselbalch, 2021).

While the principle of responsibility is common to many AI ethics documents, it tends to be poorly defined and often includes abstract discussions that pose ontological questions, such as whether AI can itself be a legal subject and therefore held accountable for automated decisions (Jobin, et al., 2019). These abstractions arguably take a highly disembodied and objective view of machine learning that prevents the proper assignment of responsibility to real decision-makers. Indeed, Terzis (2020) argues that AI ethics is “a genuine fallacy” on the grounds that it fails to address “one of the most fundamental issues in the field of advanced computational technology: the freedom and the subjectivity of all the agents involved, be it the CEO of a tech-giant, the project manager, the business analyst, the developer or the micro-worker” [24]. As we have argued, achieving and maintaining feminist data ethics of care requires an equitable distribution of work to all stakeholders in the machine learning economy, from micro-workers to chief executive officers. When responsibility for implementing ethical principles within an organisation is not widespread, the organisation’s ethical, technical and economic objectives may be less likely to converge. For example, at Google in 2021, Timnit Gebru, a high-profile leader of an AI ethics research group within the company, was asked by Google executives to retract a research paper she authored that warned of problems with a new AI system (Simonite, 2021). Google was, at the time of this controversy, using the system as part of its search engine (Simonite, 2021). Gebru ultimately left Google with both parties embroiled in a high-profile public dispute and seemingly little to no reconciliation of the ethical and commercial imperatives at stake.

The propensity for abstraction over widespread and grounded responsibility is evident in the principle of nonmaleficence which is also common to AI ethics content. The concept of nonmaleficence is typically used to articulate the idea that AI should not cause harm, and to raise concerns for individuals’ safety, security, and privacy, and potential discrimination (Jobin, et al., 2019). The problem is that concepts of discrimination and harm tend to be presented as abstractions and detached from real actors, or actions, within the machine learning economy (Terzis, 2020). In particular, the question of who is causing harm is rarely addressed and so responsibility for harm prevention is rarely properly assigned. As we contend in our feminist data ethics of care praxis, harm prevention should be intersectional, requiring a multifaceted and systemic approach grounded in the lived experiences of both the subjects of machine learning technologies and individuals working in the machine learning economy.
In AI ethics content, calls for justice and fairness are typically framed around the prevention of bias and discrimination in the use of AI-enabled systems. The values of diversity, inclusion and equality are often evoked in these discussions, yet they tend to be conceived narrowly as individual rights. For example, AI ethics content often calls for the protection of individual rights to due process, explanation, appeal, redress and remedy (Jobin, et al., 2019). While these rights are supportive of a democratic system of governance, a limitation of rights-based approaches to ethics is that they can be founded on an erroneous assumption that all people are equally capable of exercising rights, while failing to consider limits to accessing or asserting rights that stem from social, economic and other intersectional inequities (Miller and Redhead, 2019).

Similarly, a rights-based approach to justice risks placing the responsibility for responding to harm (a form of repair work) onto those subjected to harms (Kapoor, 2019). It can also lead to a focus on “discrete ‘bad actors’” and on individual cases of unfairness or discrimination rather than a wholistic view of systemic inequality that can arise from a confluence of sources operating over time. For example, a rights-based framework arguably offers little to those who may wish to prevent a machine learning enabled word association system linking “homemaker” with women and “computer programmer” with men (Bolukbasi, et al., 2016). In practice, the existence of rights alone is not sufficient to ensure just and equitable outcomes in the machine learning economy.

To undertake a feminist ethics of care approach to justice, once again we argue that the responsibility for ensuring fair outcomes must be assigned to real people and real organisations at all stages in the machine learning pipeline. Agency should not be assigned to a disembodied machine but to specific people and organisations in various parts of the machine learning economy. Ensuring diverse representation across the machine learning workforce, evaluating positionality, engaging with communities and individuals throughout the design process, and the assignment of specific responsibilities to real people, are all necessary for feminist data ethics of care outcomes. When this does not occur, and performance of a model or technical system is the only evaluative criteria of concern, it can lead to “a focus on short-term quantities (at the expense of long-term concerns)” [26]. For example, “short term incentives” (e.g., quarterly earnings) have arguably led to YouTube’s algorithms promoting white supremacy and other hateful content at various points in time (Riberio, et al., 2019). Our feminist data ethics of care framework requires thoughtful engagement with the spectrum of machine learning practices and widely distributing the responsibility for harm mitigation.

The focus on individual rights and tendency for abstraction in mainstream AI ethics content might be explained by a dominance of male voices in AI discourses. As Hagendorff (2020) describes, “no different from other parts of AI research, the discourse on AI ethics is also primarily shaped by men” [28]. This means that the principles embodied in AI ethics content are often written from and reinforce a patriarchal viewpoint, and are rarely informed by intersectional understandings of power imbalances and experiences of systemic discrimination. Interestingly, Hagendorff’s (2020) meta-analysis of AI ethics content found that content authored by women-led organisations was more likely to present AI “within a larger network of social and ecological dependencies and relationships ... corresponding most closely with the ideas and tenets of an ethics of care” [29]. In contrast, some male authored AI ethics content tends to treat ethical issues as technical problems in need of technical solutions, which perhaps explains why Google’s main solution to its machine vision algorithm classifying black people as gorillas was blocking the image recognition system from identifying gorillas (Vincent, 2018). In Hagendorff’s study, male authored ethical frameworks tend to be disconnected from their wider social and political context and failed to address “AI in context of care, nurture, help, welfare, social responsibility or ecological networks” — that is, the socio-technical nature of AI.

Of course, abstraction in AI ethics content may also arise through necessity. When seeking to capture a range of contested ideas about complex socio-technical systems into a neat organising statement or list, specificity is likely to lose out to abstraction (Terzis, 2020). However, this speaks to the limitations of such an approach to AI ethics, rather than to the merits of abstraction. Abstraction is a fundamental problem for the field of machine learning ethics because it limits the potential for assigning responsibility to real actors. To have effect, the field AI ethics must move beyond organising statements, and the responsibility for addressing and preventing harm must be more equitably distributed. Under current conditions, in which the
A feminist data ethics of care framework for machine learning: The what, why, who and how

...field of data science is dominated by a heteropatriarchal class of men (see generally Chang, 2019), a more equitable distribution of the responsibility for operationalising a feminist data ethics of care praxis in the machine learning economy depends on everyone — women, men, and non-binary and transgender people — undertaking more ethics work in their day-to-day practices. To this end, as we outline next, there are a range of solutions and policies that are available now for stakeholders to implement.

4. How: Operationalising feminist ethics of care for machine learning

Thus far we have argued that a missing link between AI ethics principles and applied AI ethics is assigned responsibility to real actors within the machine learning economy. In this Section, we aim to situate our feminist data ethics of care praxis firmly within the current machine learning economy, and to outline the range of technical tools and systems that are currently available for operationalising feminist data ethics of care (Hagendorff, 2020). We ultimately sketch a practical guide for matching existing methods, those underpinned by the principles in our praxis, with specific organisations and actors.

The first step towards achieving a feminist data ethics of care in practice is reconceptualising AI as an industry rather than an amorphous abstraction. Terzis (2020) explains:

There is no ‘ethical AI’, unless there is an ‘ethical’ supply chain, clean from conflict materials; an ‘ethical’ UX/UI design, free from manipulative dark patterns; an ‘ethical’ workforce, comprised of ethical micro-workers; an ‘ethical’ decision-making process, conditioned on the parameter of climate justice; and eventually, an ‘ethical’ reconsideration of the ‘data-driven’ necessity. [31]

The practice of ethics, in a broad sense, requires that organisations that provide machine learning infrastructure, and individuals who participate in the development and deployment of machine learning technologies, are held responsible for the individual and societal impacts of their industry. As Neff (2020) warns, the globally distributed nature of the machine learning economy which features “long global supply chains of AI systems — from data labelling work to engineering work to the front-line use of dashboard systems” can “mask the opportunities that people have for intervening in the systems, making it hard for people to contest their results, and blur the lines of accountability and responsibility” [32]. To help demystify this economy, in Table 1, we suggest the machine learning economy can be divided into three broad sectors: machine learning infrastructure, machine learning development and machine learning deployment. These sectors are not mutually exclusive or exhaustive, but the divisions are helpful for identifying and assigning responsibility to different actors within the economy and, ultimately, the different actions they might undertake to operationalise a feminist data ethics of care. This taxonomy offers a starting point for identifying the “social, structural and institutional configurations that enable and constrain individual action” in order to identify “pathways for influencing the design, application, modification and use of responsible AI technologies” [33].

| Table 1: Who and what in the machine learning economy. |
|-----------|----------------|----------------|
| **Infrastructure** | **Development** | **Deployment** |
| Infrastructure | Data analytics, | End users |
|                |                 |               |
A feminist data ethics of care framework for machine learning: The what, why, who and how

<table>
<thead>
<tr>
<th>WHO is involved?</th>
<th>WHAT are they doing?</th>
</tr>
</thead>
<tbody>
<tr>
<td>providers, e.g., Google, Amazon, OpenAI</td>
<td>Providing machine learning infrastructure including pre-trained models and cloud storage</td>
</tr>
<tr>
<td>software development, consultants, and business services</td>
<td>Providing machine learning “solutions” to businesses and governments</td>
</tr>
<tr>
<td>(organisations and individuals)</td>
<td>Using ML enabled tools and systems for automated decision-making and/or to gain insights for human decision-making</td>
</tr>
<tr>
<td>Data scientists, data engineers, company executives, micro-workers</td>
<td>Private companies, government agencies, universities</td>
</tr>
</tbody>
</table>

The second step towards a feminist ethics of care in practice is a change in consciousness of participants in the machine learning economy that empowers them to be agents of action. For data scientists, engineers, software developers and company executives to perceive themselves responsible for feminist data ethics of care work, they must first perceive themselves as autonomous decision-makers. This requires a “business ecosystem where individuals will enjoy a zone of autonomy and will be incentivized to understand and embrace full responsibility” [34]. Terzis (2020) explains:

an alternative framework for building ethical AI would be founded on the subjectivity of the individuals involved in the development process. This framework, free from ‘objectively’ ethical criteria, will take the form of a reflective process during which the individuals will obtain a clear understanding of their freedom and responsibility. [35]

Systemic ethical practice requires action from autonomous individuals working within the machine learning economy who share a common social consciousness and responsibility for preventing harm to everyday users. This will require organisational cultures that are supportive of equitably distributed responsibility for operationalising a feminist data ethics of care praxis. Khurana (2020) suggests giving workers the “opportunity to develop the right habits and frames of mind” [36] through regular workplace exercises that involve complex ethical problem solving, unstructured time for reflection, and diverse interdisciplinary work environments. Another intervention is what Vakkuri, et al. (2020) call an “ethical card deck” in which “cards pose questions to the developers and answering these questions necessitates ethical consideration from the developers. Using the cards produces transparency by producing documentation, especially related to the development process” [37]. New roles such as chief social work officer may be used to lead the integration of “social work thinking” into processes of machine learning development [38]. This might involve community consultation and impact assessments for better contextualising machine learning technologies (Patton, 2020).
The inclusion of communities affected by machine learning through participatory and co-design methods (Katell, et al., 2020) is another straightforward practical intervention available to participants in the machine learning economy. As we have argued, community engagement is important for achieving outcomes that “address problems in their situated context and re-centre power with those most disparately affected by the harms of algorithmic systems” [39]. Importantly, to ensure an equitable distribution of feminist data ethics of care work, community engagement must be commonplace and widespread within an organisation and not relegated to teams working exclusively on ethical problems. They must be integrated into design and deployment projects from the very beginning.

Once real people working within the machine learning economy are empowered, or motivated, to be responsible decision-makers, they have available to them a range of technical solutions for implementing feminist data ethics of care. As Morley, et al. (2020) point out: “how to apply ethics to the development of machine learning is an open question that can be solved in a multitude of different ways at different scales and in different contexts” [40]. Following Morley, et al. (2020), rather than attempting to assert objective and static standards for undertaking a feminist data ethics of care praxis, we argue there should always be a focus on mixed methods. People working within the machine learning economy should take up a range of solutions that are available to them now, but they should also undertake active reflection and iteration, always seeking to change culture and build capacity throughout the machine learning economy.

More specifically, there are toolkits for detecting, mitigating and auditing discrimination and bias, such as FAT ML, XAI community, AI Fairness 360 toolkit, What-If Tool, Facets, fairlern.py and Fairness Flow. Another is what Gebru, et al. (2020) call “datasheets for datasets” for documenting the “motivation, composition, collection process, recommended uses, and so on” of datasets used in the machine learning economy. Datasheets can be “implemented easily and concretely” [41], help to address the lack of widely applied industry standards for documenting machine learning datasets, and potentially improve “transparency and accountability within the machine learning community” [42]. Gebru, et al. (2020) argue that datasheets also have the potential to “mitigate unwanted biases in machine learning systems, facilitate greater reproducibility of machine learning results, and help researchers and practitioners select more appropriate datasets for their chosen tasks” [43]. As datasheets are meant be completed by real people (and not automated systems), they might also provide an opportunity for teams to undertake and document positionality and community engagement, with a view to anticipating potential biases and preventing harmful outcomes.

Overall, to undertake our praxis, a broad range of actors within the machine learning economy must be enabled to develop a social consciousness supportive of feminist data ethics of care work (Poursabzi-Sangdeh, et al., 2018). In Table 2, we map our feminist data ethics of care framework onto the sectors of the machine learning economy identified above, and we connect them to methods for operationalising the components of our praxis. This roadmap shows that a mix of technical, social and cultural interventions may be combined to enact a feminist data ethics of care in practice, across the machine learning economy at large.

<table>
<thead>
<tr>
<th>WHO is involved?</th>
<th>Infrastructure</th>
<th>Development</th>
<th>Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infrastructure providers, e.g., Google, Amazon, OpenAI</td>
<td>Data analytics, software development, consultants, and business services</td>
<td>End users (organisations and individuals)</td>
<td></td>
</tr>
</tbody>
</table>
### A feminist data ethics of care framework for machine learning: The what, why, who and how

<table>
<thead>
<tr>
<th>WHAT are they doing?</th>
<th>Data scientists, data engineers, company executives, micro-workers</th>
<th>Data scientists, software developers, company executives</th>
<th>Private companies, government agencies, universities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Providing machine learning infrastructure including pre-trained models and cloud storage</td>
<td>Providing machine learning “solutions” to businesses and governments</td>
<td>Using ML enabled tools and systems for automated decision-making and/or to gain insights for human decision-making</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HOW can they enact a feminist data ethics of care for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity within teams and in executive positions through recruitment, training, and other measures to address the ‘leaky pipeline’</td>
</tr>
<tr>
<td>Regular workplace ethics exercises including positionality evaluation and documentation, and unstructured time for reflection</td>
</tr>
<tr>
<td>Chief Social</td>
</tr>
<tr>
<td>PRINCIPLES</td>
</tr>
<tr>
<td>(i) ensuring diverse representation and participation</td>
</tr>
<tr>
<td>(ii) critically evaluating positionality</td>
</tr>
<tr>
<td>(iii) centring human subjects at every stage of the pipeline</td>
</tr>
<tr>
<td>(iv) implementing iterative transparency and accountability measures</td>
</tr>
<tr>
<td>(v) equitable distribution of care work</td>
</tr>
<tr>
<td>PRACTICES</td>
</tr>
<tr>
<td>Diversity within teams and in executive positions through recruitment, training, and other measures to address the ‘leaky pipeline’</td>
</tr>
<tr>
<td>Community consultation and participation in data collection and system design</td>
</tr>
<tr>
<td>Bias and discrimination detection and</td>
</tr>
<tr>
<td>Diversity within teams and leadership</td>
</tr>
<tr>
<td>Community consultation and participation in data collection and system design</td>
</tr>
<tr>
<td><strong>Machine Learning</strong>?</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Datasheets and other reporting standards aimed at documenting decisions, anticipating biases and preventing harmful outcomes</td>
</tr>
</tbody>
</table>

### 5. Conclusion and future research

This article conceptualised and provided an initial roadmap for operationalising a feminist data ethics of care for the machine learning economy. After evaluating select AI ethics content against our feminist data ethics of care framework, we identified a propensity for abstraction over grounded responsibility that often leaves the question of who is causing harm unaddressed and the responsibility for harm prevention unassigned. This is concerning given that in mainstream AI ethics conversations, limited attention has been paid to the significance of gender power relations, including the often-gendered nature of “care work” and how it might correlate with an unequal distribution of the responsibility for operationalising ethical principles in the machine learning economy. For those seeking to address this problem, we identified a range of practical methods and other interventions that are available to operationalise, at least in part, our feminist data ethics of care framework.

The theory and practice of feminist data ethics of care for machine learning provides fertile ground for future research. There is a particular need for work that further develops the ‘how’ of operationalising a feminist data ethics of care, including built-for-purpose tools, methods and techniques for all stages of the
A feminist data ethics of care framework for machine learning: The what, why, who and how

machine learning pipeline. Given that this technology is constantly evolving, and its applications change over time, there is also a need for research that conceptualises and measures harm. When attempting to prevent and address intersectional harms through a feminist data ethics of care framework, an equitable distribution of responsibility for this work will be critical in all contexts, from machine learning to the AI economy more broadly.

About the authors

Dr. Joanne Gray is a Lecturer in the School of Communication at the Queensland University of Technology and a Chief Investigator at the Digital Media Research Centre. Her research focuses on platform policy and governance, including the exercise of private power through automated technologies.
E-mail: joanne [dot] e [dot] gray [at] qut [dot] edu [dot] au

Dr. Alice Witt is a Postdoctoral Research Fellow in the Digital Media Research Centre, Faculty of Business and Law at the Queensland University of Technology. Her research investigates the exercise of governing power in the digital age, focusing on the intersections of regulation, technology, and gender.
E-mail: ae [dot] witt [at] qut [dot] edu [dot] au

Acknowledgements

The authors wish to acknowledge and express thanks to their colleagues at the Digital Media Research Centre, particularly Professor Dan Angus and the members of the feminist machine learning working group, who provided valuable feedback on earlier drafts of the paper. In support of this project, Dr Gray received funding from QUT’s Women in Research committee.

Notes

2. A.E. Marwick, 2019, p. 310.
4. Our feminist agenda is inclusive of the interests of womxn; that is, women defined broadly to include those who identify as cis-gendered, transgender and non-binary.
24. Terzis, 2020, p. 221.
27. Thomas and Umisky, 2020, p. 4.
32. Neff, 2020, p. 5.
33. Ibid.
34. Terzis, 2020, p. 226.
35. Terzis, 2020, p. 228.
38. Patton, 2020, p. 86.
A feminist data ethics of care framework for machine learning: The what, why, who and how

41. Hagendorff, 2020, p. 111.
42. Gebru, et al., 2020, p. 2.
43. Ibid.

References


A feminist data ethics of care framework for machine learning: The what, why, who and how


A feminist data ethics of care framework for machine learning: The what, why, who and how


A feminist data ethics of care framework for machine learning: The what, why, who and how


A feminist data ethics of care framework for machine learning: The what, why, who and how


Editorial history

Received 30 July 2021; revised 4 October 2021; accepted 15 November 2021.

This paper is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

A feminist data ethics of care framework for machine learning: The what, why, who and how by Joanne E. Gray and Alice Witt.
First Monday, volume 26, number 12 (December 2021).
doi: https://dx.doi.org/10.5210/fm.v26i12.11833