Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race

Susan Leavy
Insight Centre for Data Analytics
School of Information and Communication Studies
University College Dublin
Dublin, Ireland
susan.leavy@ucd.ie

Eugenia Siapera
School of Information and Communication Studies
University College Dublin
Dublin, Ireland
eugenia.siapera@cs.ucd.ie

Barry O'Sullivan
Insight Centre for Data Analytics
School of Computer Science and Information Technology
University College Cork
Cork, Ireland
b.osullivan@cs.ucc.ie

ABSTRACT

The potential for bias embedded in data to lead to the perpetuation of social injustice though Artificial Intelligence (AI) necessitates an urgent reform of data curation practices for AI systems, especially those based on machine learning. Without appropriate ethical and regulatory frameworks there is a risk that decades of advances in human rights and civil liberties may be undermined. This paper proposes an approach to data curation for AI, grounded in feminist epistemology and informed by critical theories of race and feminist principles. The objective of this approach is to support critical evaluation of the social dynamics of power embedded in data for AI systems. We propose a set of fundamental guiding principles for ethical data curation that address the social construction of knowledge, call for inclusion of subjugated and new forms of knowledge, support critical evaluation of theoretical concepts within data and recognise the reflexive nature of knowledge. In developing this ethical framework for data curation, we aim to contribute to a virtue ethics for AI and ensure protection of fundamental and human rights.

KEYWORDS

Ethical AI, data curation, feminist theory, critical theories of race

ACM Reference Format:
Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. 2021. Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3461702.3462598

1 INTRODUCTION

Artificial Intelligence systems that have been found to discriminate, have invariably disadvantaged those already most marginalised in society. This follows a long history of injustice resulting from bias in the representation, classification and categorisation of people in data [30, 55, 71]. While the very possibility of attaining value-free, objective knowledge and representing this in data has long been debated (see [28, 37]), the availability of large data sets and machine learning promised a new data-driven empiricism that, given sufficient data, could uncover objective truth. However, mounting evidence demonstrating the capacity for machine learning algorithms to learn and often amplify a range of societal biases, exposed fundamental issues with this approach. There is now, widespread acknowledgement of bias as an *"an unavoidable characteristic of data collected from human processes"* [17].

Cases of discrimination in AI systems have been traced to a large extent to human decisions concerning how people are represented in data and the creation, collection and annotation of datasets. As Bartoletti proposed, *"it is time we acknowledge that data is simply not neutral, and, as such, every single decision and action around data is a political one"* [2]. Understanding decisions involved in data curation for AI as political necessitates an entire reform of the process and questions the very possibility of developing ideologically neutral AI systems. However, even if ideological neutrality may be illusive, bias in data, particularly in relation to groups of people, implies an expression of negative sentiment or representations that are prejudicial or disadvantageous to them and thus, curating data in a way that identifies bias and insists on fairness is feasible.

Approaches to ensuring fairness in AI have broadly taken a normative approach through the development of ethical guidelines and principles [27, 73]. Given the level of abstraction inherent in such guidelines, there is also a need to promote the development of a virtue ethics among technology developers [41, 67]. The objective of this paper is to develop a critical framework for practitioners involved in each stage of the process of developing AI systems to enable critical reflection and responsibility for the potential effects of the use of data.

In this paper we translate feminist epistemology along with principles of social justice and critical theories of race and gender into a a framework for the ethical curation of data for AI systems with the aim of bringing to light underlying power structures embedded in training data for machine learning. We focus primarily on discrimination on the basis of race and gender, while acknowledging that discrimination may also occur based on a range of other factors and such as socio-economic background, religion or levels of ability along with intersections of such identities.

AIES '21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462598

2 DISCRIMINATION IN AI

Central tenets underlying human rights and civil rights movements concern attaining rights to self-determination and freedom from being classified, stereotyped and treated differently. The profound consequences of issues with the design of AI systems and their capacity to perpetuate discrimination have resurfaced decades-old debates concerning social justice and the surveillance of racialized and gendered bodies. Cases of race and gender discrimination due to the way individuals were categorised and treated differently by AI algorithms have been uncovered, demonstrating the capacity of AI to mirror and even exacerbate the discriminatory behaviour that civil rights movements have fought against [15, 38, 51]. Such cases highlight the fundamental ethical problems with using AI to automatically classify individuals into groups, generalising based on assumed identities and subsequently treating them differently based on such categorisations.

The automation of decisions using AI algorithms has also highlighted discriminatory practices within existing human processes. For instance, Noble [51] uncovered a series of cases where data-driven racial profiling of individuals resulted in the repetition of historical injustices against people of colour through what she termed *"technological redlining"*. Such classification of people online is often based on assumed attributes or *"affinity profiling"*. Wachter [70] argues that categorising and treating people differently in this way fundamentally undermines the right to privacy and non-discrimination and proposes strategies to address this risk of proxy discrimination.

The danger presented by *"runaway feedback loops"* [23] generated by machine learning algorithms were highlighted in the context of the Black Lives Matter protests. Predictive policing systems trained on historical police records are being increasingly used to forecast criminal behaviour despite the accuracy of this data having been called into question and the evident discrepancy between the real occurrence of crime and how it is recorded [45, 56]. A predictive machine learning algorithm commonly used in the US to support healthcare decisions was also found to produce racially biased decisions [52]. The system used historical data concerning healthcare spend to predict future healthcare need and recommend decisions accordingly. This overlooked how racial differences in historical healthcare spend were attributable to disparities in wealth rather than a real reflection of healthcare need, thus perpetuating patterns of historical social injustice.

Language in particular, is imbued with stereotypical societal concepts and unravelling such biases from text data can be particularly challenging . The extent to which historical patterns of discrimination in society can be learned by neural embedding techniques was demonstrated in an analysis of a model trained on text covering 100 years dating from 1910 [22]. The learned associations however, reflected employment figures in the US at the time. In this sense, the model may be considered descriptive of a particular time rather than inherently biased. However, the text represented an unjust societal system and in that sense reflects the biases of that time. This demonstrates how, through using historical data and learning concepts that would now be considered biased or stereotypical, historical injustice may be reinforced. In fact, many studies in digital humanities use machine learning techniques precisely because they are effective in modelling biases and uncovering historical concepts and inequalities embedded in language (eg. [75]). Such bias in language models is evident in some translation systems and co-reference resolution algorithms [68, 76]. Image data sets have also been shown to lead to biased algorithms [62]. Facial recognition systems were found to be less accurate for those with darker skin and females [7]. However, despite demonstrable racial bias in the applications, technology such as facial recognition is becoming increasingly integrated with core security and border control infrastructures. Given the evident threat to social justice and fundamental human rights posed by AI systems learning from data, it is clear that a substantial change is required to methods for AI data curation.

3 THE PROBLEM WITH FAIRNESS AND BIAS

A considerable challenge in dealing with discrimination and bias in AI generally, and in machine learning in particular, is differing definitions of fairness and bias. According to Selbest et al. [57] the issues are in large part due to an abstraction of the concept of fairness away from the social context within which they are to be deployed. We consider here the often discussed COMPAS system for predicting recidivism.[1] Tables 1 and 2 are based on an analysis presented by Sumpter [63] that considered whether or not COMPAS exhibits racial bias. The answer is that it depends on what one defines as bias.

Kleinberg et al. [36] considered three different fairness conditions that capture what it means for an algorithm to be fair and avoid exhibiting bias. The first of these is that the algorithm should be *well calibrated*. Essentially, this means that the proportion of people that are, for example, classified as positive in a population should be equal to the proportion of people that are positive in each subgroup of the population. Comparing Tables 1 and 2 we see the that $1369/2174 = 63\%$ of African American defendants classified as high risk, while $505/854 = 59.1\%$ of White defendants were classified as such. This data is well-calibrated, suggesting the system does not exhibit a racial bias, since the proportion of high-risk people in each population is predicted to be roughly equal.

The second (and third) condition relates to balancing for the positive (resp. negative) class. If this condition were to be violated it would mean that the likelihood that a positive (resp. negative) instance in one population is more likely to be identified than in the other. For example, as Sumpter argues, Tables 1 and 2 show that $2174/3615 = 60\%$ of African Americans in COMPAS are considered higher risk, while this only $854/2464 = 35\%$ for Whites; note that $1901/3615 = 53\%$ of African Americans reoffended while this was $966/2454 = 39\%$ for White, suggesting that African Americans actually reoffended less than predicted while the opposite is true for Whites. This suggests a strong racial bias.

Considering the mistakes that COMPAS makes for each racial group, Sumpter goes on to show that $805/1714 = 47\%$ of African Americans were wrongly predicted to reoffend, as compared with only $349/1488 = 24\%$ of Whites. On the other hand, only $532/1901 = 28\%$ of African Americans who reoffended were wrongly predicted

[1] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Table 1: Recidivism rates in the COMPAS system for African American defendants (via [Sumpter, 2018]).

African American	High Risk	Low Risk	Total
Reoffended	1,369	532	1,901
Didn't reoffend	805	990	1,714
Total	2,174	1,522	3,615

Table 2: Recidivism rates in the COMPAS system for White defendents (via [Sumpter, 2018]).

White	High Risk	Low Risk	Total
Reoffended	505	461	966
Didn't reoffend	349	1,139	1,488
Total	854	1,600	2,454

as being lower risk, as compared with 461/966 = 48% of Whites. Again, this suggests a strong racial bias.

In other words, while COMPAS satisfies one of Kleinberg et al.'s definition of fairness (calibration), is fails the second and third. This is not unusual. In fact, Kleinberg et al. rigorously prove that these three fairness conditions are incompatible except in very rare situations. This implies that one is usually faced with a choice of which bias must be traded-off against another. Eliminating bias from AI systems is often provably impossible. Furthermore, even if all three conditions were met by COMPAS, the tool's predictions are based on a model of policing that currently being criticised as institutionally racist [45, 56]. In this manner, we cannot separate AI systems from the wider social context.

4 MITIGATING BIAS AND DISCRIMINATION IN AI

AI systems are imbued with the values of those who design, develop and commission technology, as evidenced by the fact that privileged groups in society have not generally been subject to algorithmic discrimination [50, 74]. To develop a self-reflective critical practice among AI developers a critical science approach founded in decolonial theory was proposed by Mohamed et al. [48]. The authors proposed that decolonial studies along with critical theories of race, feminism, law, queerness and science and technology studies form the basis of a *"self-reflexive approach to developing and deploying AI that recognises power imbalances and its implicit value-system"*. To counteract the imposition of the standpoint of the designer on technology, Constanza-Chock [12] proposed *"design justice"* which outlines a participatory design process that makes a consideration of diverse standpoints through necessary. A comprehensive critique founded in intersectional feminism of the power dynamics at play in who records data and who is recorded is detailed by D'Ignazio and Klein [16].

In addressing imbalances in data resulting from an under representation of particular groups, one response has been to collect more data from that group of people. This has led to questions of unethical data gathering practices with the goal of balancing datasets [29]. Benjamin [3] examines the complex socio-political ramifications of this in the context of race, questioning what it means *"to be included, and hence more accurately identifiable, in an unjust set of social relations?"*. It may seem fair to increase the inclusion of people of colour in datasets to improve accuracy of facial recognition systems. However this may be in a broader context where such systems may be used to perpetuate social injustice and exacerbate existing social inequalities. The inequitable treatment of the poorer in society in relation to data collection is described by Eubanks [18] as the *'Digital Poorhouse'*, where more data is collected about disadvantaged groups and that this data ends up being used in ways that perpetuate social and economic inequality.

Technical approaches to the prevention of bias and discrimination in AI have included developing fairness-aware machine learning algorithms, modifying learned models along with addressing bias in data. Methods that address bias in data include methods for data augmentation, re-sampling, re-weighting, swapping labels and removing dependencies between characteristics with the aim of achieving neutrality [19, 34]. Within narrowly defined contexts, modifying data to reduce bias has been shown to reduce stereotypical associations within algorithms and improve fairness in outcomes. In language technology for instance, binary categories of gender mentioned in text data was swapped [54, 76]. Other work focused on amending learned models and disassociating stereotypical associations between entities in word embedding models [4, 10, 64]. However, given that judgements as to what is biased and what constitutes a fair representation of a person's identity or community is based on a particular standpoint, principle or context, it follows that each attempt to reduce bias is in practice aligning it with a different set of principles or standpoint. As part of ensuring the development of ethical AI it is important therefore that the philosophical standpoint undertaken is made explicit and transparent.

5 THEORETICAL GROUNDING FOR ETHICAL DATA CURATION

6 FEMINIST EPISTEMOLOGY

This paper proposes a critical framework for the curation of data for AI with a focus on evaluating theoretical standpoints and social concepts that may underlie a data set. To achieve this, the framework is grounded in epistemological assumptions shared by post-structuralist and intersectional feminist theories, as outlined by Mann [46]. We integrate further insights from critical theories of race, intersectional feminism and decolonial theory, to develop actionable principles for the curation of machine learning data [14, 28, 69]. Core assumptions concern a social constructionist view of knowledge, a call for excavation and retrieval of subjugated knowledges, an exploration of new sites and forms of theorising and an understanding of the nature of knowledge as reflexive.

6.1 Critical Theories of Race and Feminist Principles

Many of the debates concerning artificial intelligence and discrimination, echo decades of scholarship and activism concerning the representation, stereotyping and exclusion of people, particularly in the language of literature, media and public discourse [9, 21, 43]. The same critical perspectives and activism that achieved advances in social justice are fundamental to ensuring the development of

fair and ethical artificial intelligence. This was demonstrated by D'Ignazio and Klein [16] who outlined principles for a feminist approach to data science that included calls to examine and challenge power, elevate emotion and embodiment, rethink binaries and hierarchies, embrace pluralism, consider context and make labour visible. Within the specific context of machine learning, even if data used aligns with these principles, they may still generate damaging learned concepts of gender [39]. Thus we draw upon the work of D'Ignazio and Klein [16] and examine how feminist principles may be incorporated into a critical framework for ethical data curation for AI that examines data within the context of a particular machine learning approach and along with with the raw data itself, examines approaches to sampling, feature selection and data annotation all of which have the capacity to influence and skew data.

If we accept that training datasets need to be curated in order to interrogate underlying values and perspectives and also therefore to address racist, misogynist and other discriminatory contents, then the question that arises is: how do we recognise these contents? A central premise of race critical approaches is that racism is adaptable and a scavenging ideology [61]. It borrows ideas and concepts from elsewhere, it makes use of different frames and metaphors, and it often functions through constant debating and denial [40, 66]. Denial of racism is very common and there is little certainty on what is racist, with the exception of clear statements of inferiority and incitement to violence and hated, which form part of what is in Europe illegal hate speech [53]. Beyond illegal hate speech there is little agreement on which kinds of contents are racist, misogynist, or contain other forms of social hate.

The prevention of racial discrimination in algorithms demands a move from the normative ideal of colour blindness towards a stance informed by race critical theories (e.g. Essed and Goldberg 2002) for instance, that seek to identify the ways in which current institutions, such as the law, may embody white supremacy and actively oppress people of colour. Lentin [40] understands the construct of race as a kind of technology that seeks to separate and subjugate black people and people of colour thereby perpetuating white supremacy; racism, in other words, presupposes race. This however does not mean that 'race' can be removed easily, because it is integral to many of the systems of modern society, for example, law and law enforcement, education, health, employment and so on. Remedial measures therefore include firstly, the identification of the *"apparatus of oppression"* including the ways in which institutions are used to control and subjugate people of colour and the cultural assumptions around 'race'. Secondly, and following from the first strategy, the development of an arsenal of tools that dismantle white supremacy and redress racial injustices.

It is crucial to note here that these strategies require the collaboration of various kinds of knowledge, including the academic and specialist knowledge of race critical theorists, the situated knowledge of those who are living and experiencing the effects of race in their everyday lives, and the knowledge generated in the struggle to dismantle white supremacy, gained in the process of organised and quasi organised social movements, such as the Black Lives Matter movement and other anti-racist initiatives. Importing these ideas of critical race into a framework for the curation of data for AI therefore requires an active anti-racist stance. This means that assuming a colour blind approach will in fact hide rather than expose racism. We must therefore in the first instance interrogate datasets with a view to identifying embedded assumptions about race and ethnicity. Secondly, mobilise the various kinds of knowledge which can contribute to making the dataset and the system it feeds into part of the arsenal by which white supremacy can be dismantled. This will necessitate a participatory or co-design approach, which draws upon Constanza-Chock's [12] approach but also builds upon other work, for example Katell et al. [35]. These approaches propose an active engagement with communities and activist groups involved in anti-racist work. In translating epistemological assumptions to a critical evaluation framework for machine learning data we also call upon work by Gillborn et al. [25] who developed Quantitative Critical Race Theory to support analysis of statistical data. The principles devised in this work both align with feminist epistemological assumptions and propose practical insights into how data may serve an anti-racist agenda. The principles the authors noted assert the centrality of racism, that numbers are not neutral, categories are neither natural nor given, that data cannot speak for itself and finally, that quantitative data should play an active role in highlighting and combating racism.

7 ETHICAL FRAMEWORK FOR DATA CURATION FOR AI

The approach to ethical curation of data set out in this paper is grounded in feminist epistemology to enable the interrogation of theoretical standpoints and concepts underlying data for AI systems. Principles of feminist theory and critical race theories are drawn upon to formulate an ethical framework to mitigate discrimination and bias in data for AI. The framework we set out may be employed within the context of the data curation process set out by Jo and Gebru [33], who proposed a new specialisation in data curation for AI based on archival methods. The following presents a summary of the principles we set out:

(1) **Examine perspectives in data.** The perspectives of individuals involved in both data creation and the development of AI systems are embedded in curated data collections and play a central role in the social construction of concepts. Aligning with principles of feminism and critical race theory, to prevent damaging social constructs being learned by AI systems, the first point of critical examination in ethical data curation is to understand who's perspectives are encoded in data and the potential implications of this.
(2) **Recognise the reflexive nature of knowledge.** Feminist epistemology recognises that choices made in representing knowledge play a role in generating societal concepts. Aligning with feminist principles and critical race theory, an activist stance in the curation of data for AI is called upon.
(3) **Analyse theory in data.** From the standpoint of feminist epistemology, how we represent knowledge in data is heavily value-laden and influenced by philosophical viewpoints. Uncovering the nature of theoretical concepts, particularly pertaining to identity, is therefore central to ethical data curation for AI.
(4) **Include subjugated & new forms of knowledge.** Knowledge from groups considered subordinate along with forms of knowledge outside the predominate forms of discourse

are often marginalised or omitted from what is considered legitimate forms of data and can therefore be more difficult to access. To address such imbalances, ethical data curation requires inclusion of multiple sources and forms of knowledge.

7.1 Examine Perspectives in Data

From the viewpoint of feminist epistemology, knowledge represented in data is constructed from particular perspectives in society and influenced by factors such as gender, race, class and location. Given how such perspectives can be learned by machine learning algorithms, an essential part of ethically curating data for AI involves understanding whose perspective is reflected in a data set. This social constructionist view of knowledge is assumed by both intersectional and post-structuralist feminism and highlights the relationship between knowledge and power [46]. As an illustration of how the subjective perspectives of those involved in data creation can be integrated into an AI system, biases of healthcare workers embedded in patient records were shown to be learned by an AI decision-support system [1].

Race as a socially constructed concept is widely accepted within social research [49] and is a central tenet of critical race theory [42]. Simone de Beauvoir cited the social construction of gender as *"the fundamental source of women's oppression"* [59]. This social construction occurs, according to Butler [8], largely through language. This highlights the potential role emerging natural language generation technologies have in influencing the social construction of gender in society. Aligning with a social constructionist view of race, Gilborn et al. [25] described how information is documented in a way that reflects the interests, assumptions and perceptions of the most powerful and can result in the *"colonization of interpretation"*. To address how racism can be reflected in data therefore, the authors recommend a critical examination of data, how it is interpreted and identify potential racist, misogynist and other forms and patterns of thinking that may be embedded within it.

A new dimension in the curation of data for machine learning is the cognitive labour of classifying and labelling data and how this is increasingly conducted by underpaid women of colour from the global south [13, 16]. While they may be extensively involved with the work of data preparation for AI systems, their perspectives are unlikely to be incorporated, which serves to support Benjamin's [3] call for the *"democratization of data"* through the inclusion of all perspectives, especially from those most vulnerable to discrimination, into the design of technology.

In operationalising concepts of social construction, D'Ignazio and Klein [16] propose a critical examination of who does the work of data science, who benefits and whose priorities are met. Given the predominant characteristics of those involved in the development of AI as outlined by the AI Now Institute [72], the dominant group whose values are likely to be captured are likely to be straight, white, cisgender, college-educated men. The process of ethical data curation for AI involves consideration of layers of perspectives incorporated into the process of refining training data including those who design the structure of data sets, create data, design labels and annotate as each layer of bias could in turn be learned and further perpetuated by a machine learning algorithm. We therefore propose the critical examination of each stage in the process of data curation of AI, considering the identities of those involved, how their perspectives and views of societal power structures may be embedded within the data and how these perspectives align with the intended application of the AI system.

The complete process of creating and curating data for AI commonly involves authoring content, sampling and selection, data representation design, feature selection, labelling and annotation. The ability to pinpoint those who are responsible for each stage in the process greatly increases its transparency and accountability. While further points in this framework set out to address strategies to address the dominance of one particular world view in data curation, the first point of critical examination is to understand whose perspectives are encoded in the data.

7.1.1 Perspectives Encoding Societal Power Structures. The importance of evaluating the perspective of the source of knowledge in data is demonstrated by critiques of Twitter's use of saliency algorithms for image cropping. In this case, women's faces and people with darker skin were cropped from images demonstrating the reinforcement of social visibility patterns whereby white men are given more prominence than black men or women. The case highlighted the need for transparency regarding who's gaze is recorded in eye-tracking experiments and whether they are representative or indeed suitable for replication in a particular context. Data as seemingly straightforward as what a person views as interesting in an image encodes many different influences that include gender and race and these perspectives must therefore be considered in the curation of training data.

Given the well documented issues with the generation of racist, sexist and otherwise toxic language in such language models (see [24]), it is crucial to know and evaluate the perspectives encoded in the data that these algorithms are trained on. For instance, the natural language generator GPT-3 launched by OpenAI was trained on samples taken from the Common Crawl, a large collection of data from the web [6]. Applying a social constructionist critique of the use of this data in this context would necessitate an examination of the perspectives encoded in the Common Crawl data set, the sources of the content, likely authors and the perspectives of the curators of the data sets.

Along with identifying the perspectives embedded in data sets, examination of how they align with the intended application of an AI system is crucial, as data that is not inherently biased may result in bias within a system if utilised within a particular context. This was demonstrated by Obermeyer et al. [52] who uncovered racial bias in an AI system used to support healthcare decisions due to the training data captured historical patterns of access to healthcare connected with wealth rather than healthcare need. Records of financial transactions associated with healthcare treatments were recorded by the hospitals, reflecting the requirements and perspectives of the charging entity. This contrasts with the objective of the AI system which was to support understanding of the healthcare needs of each individual patient, regardless of financial considerations. This case demonstrates just how data that was created from a perspective that is misaligned with the goals of an AI system can result in discriminatory decisions.

7.2 Analyse Theory in Data

From the from the standpoint of feminism or critical race theories, data is not objective, neutral or value-free. Rather, data captures aspects of reality perceived through the viewpoint of those most centrally involved in its creation. It is necessary therefore to critically examine the concepts and ideologies that underlie data for AI systems.

The importance of examining how concepts such as race is represented was captured by Gillborn et al. [25] who critiqued the encoding of race in terms of fixed observable attributes as an "approximation that risks fundamentally misunderstanding and misrepresenting the true nature of the social dynamics that are at play". Rather, to align with a perspective of critical race theory, race would be represented in data as, what Gillborn et al. termed, a *"complex, fluid and changing characteristic of society"*. For D'Ignazio and Klein [16], context is key to understanding concepts embedded in data and they call upon work by Borgman [5] who argued that data is situated within a 'knowledge infrastructure - an ecology of people, practices, technologies, institutions, material objects, and relationships'.

Within the context of AI systems, while the value-laden nature of human knowledge is generally acknowledged, often the data used is dictated by what is readily available. This can lead to a lack of critical examination of the philosophical underpinnings of key concepts in data. Within the context of race and gender for instance, certain representations can lead to the generation of proxy variables that form the basis for categorising people and treating them differently and perhaps unfairly. We therefore propose an examination from a feminist and anti-racist standpoint of theoretical conceptions of identity that are embedded in text.

7.3 Recognise the Reflexive Nature of Knowledge

Feminist epistemology assumes that knowledge is constructed by people who are subsequently constructed by them. This commonly refers to the necessity that authors acknowledge the factors that may have influenced their own knowledge and how that could in turn influence other people. The reflexive nature of knowledge production is crucial in ethical data curation for AI. How gender is produced and reproduced through knowledge generation and discourse is a central tenet of post-structuralist gender theory associated with, for example, Butler [8], who argues that gender is reproduced through performative speech acts. While Butler spoke from the context of humans constructing gender by (re)producing performative speech acts, the same dynamic of construction of gender through language could be applied to discourses that are produced by AI systems. In most AI applications data is not static but constantly being updated through addition of new data, proxy variables or automated reorganisation and re-classification of entities. In this way data is augmenting constantly based on more data being added but also on outputs from AI algorithms which are in turn transformed into data. In this way, the continuous reflexive process of knowledge creation in AI is analogous to the post-structuralist view of knowledge generation. The evolving dynamics of of how concepts of gender and race are embedded in data require therefore, a view of the process as reflexive.

Acknowledgement of the reflexive nature of knowledge compels those involved in the development of AI to recognise and adopt a theoretical standpoint in the ethical curation of AI data. Aligning with the assumption of the reflexive nature of knowledge, Gilborn et al, [25] propose that data is assessed taking an anti-racist stance conscious that race can be 'made and legitimated' through data and that this will serve to combat racism in society. Given the capacity for AI to perpetuate inequalities, we propose that an activist stance is taken in the ethical curation of data for AI and that the process reflects what Dignum [17] described as 'the world as it should be'.

7.4 Include Subjugated & New Forms of Knowledge

Feminist epistemological assumptions call for inclusion of what Foucault termed *"subjugated knowledges"* [20], referring to knowledge from groups considered subordinate, resulting in their knowledge being marginalised. This can lead to instances of vital knowledge being excluded (e.g. [32]). Often these knowledges are not documented, further exacerbating the potential for omission from data. Crucially, it can lead to the prioritisation of knowledge from a majority group that is damaging to subjugated groups. Forms of knowledge outside the dominant domain of discourse or what is considered a legitimate form is also required. This was central to Audrey Lorde's [44] critique of second wave feminism's exclusion of the knowledge and experience of black women. Applying such a principle in the context of data analysis, D'Ignazzio and Klein [16] describes how *"data feminism teaches us to value multiple forms of knowledge"*.

Critical examination of the inclusion of subjugated knowledges and new forms of knowledge and rectifying imbalances may result in gathering of data from different sources, such as for example, literature and oral histories. An alternative or complementary practice would be to solicit the knowledge and experiences of those groups that are historically marginalised and subjugated. This could be done through qualitative interviewing and a life history narrative approach. The extensive processes of digitisation of archival material, books and documents has opened up opportunities to find and retrieve a range of knowledge forms and sources.

The proliferation of social media presents an opportunity to *"follow imaginations, opinions, ideas, and feelings of hundreds of millions of people"* [47]. However, while perspectives and knowledge from a broader segment of the population may be accessible, recent trends showing how women and people of colour are leaving platforms, taking pseudonymous online identities and changing the style and content of their posts due to harassment demonstrates how voices from subjugated groups online can be marginalised [11]. Despite this, predictive models continue to be built on social media data without a full critique of the identities of those that are included or excluded from the sample. Models for instance, trained on social media data that aim to support mental healthcare decisions (e.g. [26, 65]) could, if applied to the whole population, have the effect of diverting vital health care resources away from other groups.

7.4.1 Inclusion through Participatory Design. In addressing this issue and following the main theoretical premises of race critical theory, we propose the development and deployment of a participatory design/design justice approach. Specifically, design justice

is a theoretical framework that seeks to address structural inequalities through design practices, where these practices include and prioritise forms of knowledge that come from the experiences of subjugated people [12]. Participatory design refers to practices that involve technology users in the process of designing technological systems and which seek to create technologies that are more responsive to human needs [60]. Both may be usefully mobilised in thinking about how to approach the issue of addressing forms of social hate that reinforce structures of domination such as racism and misogyny. A participatory design approach would then engage with those belonging to subjugated groups who have the necessary racial and feminist literacy to be able to identify, name and deconstruct linguistic features in the text. On the other hand, thinkers such as Benjamin [3] are critical of the term design, as it may be seen to mobilise 'design speak' to subsume long standing practices of resistance and subversion under one rubric, and may appear to want to deal with racism through innovation. Additionally, Irani and Silberman [31] argued that even in case of developing activist tools with the goal of empowering workers, designers are still given a higher status than workers, seen as 'design saviours'. Finally, in a context where subjugated people are often tasked with the additional labour of having to educate others on the processes of domination, does participatory design contribute to emancipation or does it actually contribute to further exploitation?

The goal is therefore to develop an approach that (i) is attentive to the requirement of prioritising subjugated and situated knowledge, i.e. knowledge derived from people's experiences as racialised/gendered subjects [28, 35], with the objective of contributing to the emancipation of people belonging to these (and other) subjugated groups; (ii) adopts a flat hierarchy as an organisational principle considering 'designers', consultants, and users as equivalent and necessary participants; (iii) therefore remunerates and compensates appropriately all teams members. The evaluation or curation of training datasets that have linguistic features should therefore be seen as the work of a team that should be able to handle these tasks: the task of identifying and providing examples of 'toxic' contents in race/gender and other ways and manually coding some instances (c.f. [58]) ; the technical task of using these coded instances as part of an algorithm that can identify and extract similar instances across the dataset; and the task of going through a representative sample of the extracted instances, validating them and feeding them back to sharpen the algorithm.

7.4.2 An Example of a Participatory Approach. A similar design was used in the HateTrack project [58], which sought to create a tool to score social media posts on the probability to contain racially toxic contents (0=no toxicity, 1=toxic). The team consisted of two members from a social science background and expertise in critical race theory and two members with technical skills. The former two organised interviews with members of communities targeted by racist speech, and with representatives of community-based organisations. Additionally, they collected materials that were labelled racially toxic by the interviewees. They used the interviews and the materials in order to generate a coding schedule for manually coding 600 instances of racially toxic materials and another 600 neutral references. These were then used by the technical team members to inform a neural network algorithm that formed the basis of HateTrack. The tool further had the ability to allow for manual coding of returned instances that were falsely identified as highly likely to be toxic and vice versa.

The key learning of this approach to identifying racially toxic contents is that it is crucial to enable members of targeted groups to define the terms by which they are attacked. The inclusion of the voices of those targeted made the tool much sharper in identifying instances of toxicity, and the tool was able to identify instances where the utterance in question became toxic through its association with a particular context (for example, the use of metonymies, such as 'religion of peace' which was pejoratively used to refer to Islam). However, the process of interviewing, coding and validating is labour intensive and the exposure to toxic speech is potentially damaging. Additionally, the interview process highlighted the added stress of asking those targeted by racist speech to repeat and discuss these attacks; while for some people it may have a healing effect as it enables them to share their trauma, for others, repeating these attacks traumatises them further. Finally, while people spoke to the team voluntarily and in order to develop HateTrack as a public tool, they were not compensated for their time. Ultimately, although the team members had considerable expert knowledge in race critical theory and some experience of racialisation as they were both of migrant background, most of the project knowledge on toxicity came from the interviewees who had direct experiences of racism. This resulted in a form of alienation of the interviewees as sources of knowledge from the end product, although they were all very supportive of the overall goal.

If language-based training datasets are to be evaluated or curated for racism/misogynism and other hate discourses, the inclusion of team members with specialist knowledge in race and gender theory must be considered a necessary condition. Given the mutations and context-boundedness of racially and gender-based toxic language, lived experience and situated knowledge add significant value to theory-based approaches. But, based on the experiences of the HateTrack project, and the critiques of 'design-thinking' [31], the knowledge to be used cannot be separated from those who lived and produced it. In these terms, they must be thought of as fully integrated and appropriately compensated team members. Evaluation and curation of training data for AI therefore requires the set up of a multi-disciplinary team with specialist skills and the use of different modes of knowledge production.

In deploying multi-disciplinary teams the goal is not only to evaluate and possible 'clean' training datasets from racial and gender-based toxicity; they could also contribute in building/synthesising language-based datasets from a variety of appropriate sources, including literature, life histories, first-person accounts and testimonials, and so on. In this manner, the deployment of such teams can be seen as having a dual goal: that of evaluating the health-toxicity of training datasets and in constructing synthetic datasets based on the knowledge and experiences, past and present, of members of subjugated and oppressed groups. This will lead to what we may call *'data utopianism'*: datasets will 'speak' and therefore train AI systems in the language of equality, freedom and emancipation even if these do not exist at present, addressing at least in part, Benjamin's call for new abolitionist tools.

8 CONCLUSION

The critical framework for the ethical curation of data for AI set out in this paper aims to enable an interrogation of the power structures and theoretical conceptions of identity that may be embedded in data for machine learning rather than addressing specific instances of bias. The assumptions of feminist epistemology upon which this framework is grounded along with principles of critical race theory address how the values and perspectives of those involved in the creation, collection, processing and curation of data for AI are embedded in its contents. The framework emphasises the importance of identifying groups whose knowledge may be omitted and re-balancing data accordingly. The value-laden nature of data and how this can generate bias in AI systems is examined and those involved in the development of AI systems are called upon to adopt an activist stance in the ethical curation of data. Through developing this critical framework for data curation we aim to contribute towards a virtue ethics among technology developers that would facilitate transparency and promote accountability for issues of algorithmic discrimination in AI.

REFERENCES

[1] Mehmet Eren Ahsen, Mehmet Ulvi Saygi Ayvaci, and Srinivasan Raghunathan. 2019. When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis. *Information Systems Research* 30, 1 (2019), 97–116.

[2] Ivana Bartoletti. 2020. *An Artificial Revolution: On Power, Politics and AI.* The Indigo Press.

[3] Ruha Benjamin. 2019. *Race after technology: Abolitionist tools for the new jim code.* John Wiley & Sons.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.

[5] Christine L Borgman. 2015. *Big data, little data, no data: Scholarship in the networked world.* MIT press.

[6] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[7] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[8] Judith Butler. 1988. Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre journal* 40, 4 (1988), 519–531.

[9] Judith Butler. 2011. *Gender trouble: Feminism and the subversion of identity.* routledge.

[10] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[11] Danielle J Corple and Jasmine R Linabary. 2020. From data points to people: Feminist situated ethics in online big data research. *International Journal of Social Research Methodology* 23, 2 (2020), 155–168.

[12] Sasha Costanza-Chock. 2018. Design Justice: towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society* (2018).

[13] Kate Crawford and Vladan Joler. 2018. Anatomy of an AI System. *Retrieved from: https://anatomyof.ai/* (2018).

[14] Kimberlé W Crenshaw. 2017. *On intersectionality: Essential writings.* The New Press.

[15] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies* 2015, 1 (2015), 92–112.

[16] Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism.* MIT Press.

[17] Virginia Dignum. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way.* Springer International Publishing.

[18] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

[19] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[20] Michel Foucault. 1980. *Power/knowledge: Selected interviews and other writings, 1972-1977.* Vintage.

[21] Betty Friedan. 2001. The Feminine Mystique (1963). *New York* (2001).

[22] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[23] Timnit Gebru. 2019. Oxford Handbook on AI Ethics Book Chapter on Race and Gender. *arXiv preprint arXiv:1908.06165* (2019).

[24] Sam Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *arXiv preprint arXiv:2009.11462* (2020).

[25] David Gillborn, Paul Warmington, and Sean Demack. 2018. QuantCrit: education, policy,'Big Data'and principles for a critical race theory of statistics. *Race Ethnicity and Education* 21, 2 (2018), 158–179.

[26] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43–49.

[27] Thilo Hagendorff. 2020. The ethics of Ai ethics: An evaluation of guidelines. *Minds and Machines* (2020), 1–22.

[28] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies* 14, 3 (1988), 575–599.

[29] Amy Hawkins. 2018. Beijing's big brother tech needs African faces. *Foreign Policy* 24 (2018).

[30] Christine B Hickman. 1997. The devil and the one drop rule: Racial categories, African Americans, and the US census. *Michigan Law Review* 95, 5 (1997), 1161–1265.

[31] Lilly C Irani and M Six Silberman. 2016. Stories We Tell About Labor: Turkopticon and the Trouble with" Design". In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4573–4586.

[32] Richard Jackson. 2012. Unknown knowns: The subjugated knowledge of terrorism studies. *Critical Studies on Terrorism* 5, 1 (2012), 11–29.

[33] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 306–316.

[34] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[35] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 45–55.

[36] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA (LIPIcs, Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[37] Thomas S Kuhn and Joseph Epstein. 1979. The essential tension.

[38] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science* (2019).

[39] Susan Leavy, Gerardine Meaney, Karen Wade, and Derek Greene. 2020. Mitigating Gender Bias in Machine Learning Datasets. In *Report on the International Workshop on Algorithmic Bias in Search and Recommendation (Bias 2020)*.

[40] Alana Lentin. 2018. Beyond denial:'not racism'as racist violence. *Continuum* 32, 4 (2018), 400–414.

[41] Sabina Leonelli. 2016. Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2083 (2016), 20160122.

[42] Ian F Haney Lopez. 1994. The social construction of race: Some observations on illusion, fabrication, and choice. *Harv CR-CLL Rev.* 29 (1994), 1.

[43] Audre Lorde. 1980. Age, race, class, and sex: Women redefining difference. *Women in Culture: An intersectional anthology for gender and women's studies* (1980), 16–22.

[44] Audre Lorde. 2020. *Sister outsider: Essays and speeches.* Penguin Classics.

[45] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[46] Susan Archer Mann. 2013. Third wave feminism's unhappy marriage of post-structuralism and intersectionality theory. *Journal of feminist scholarship* 4, 4 (2013), 54–73.

[47] Lev Manovich. 2011. Trending: The promises and the challenges of big social data. *Debates in the digital humanities* 2, 1 (2011), 460–475.

[48] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology* (2020), 1–26.

[49] Ann Morning. 2007. "Everyone Knows It's a Social Construct": Contemporary Science and the Nature of Race. *Sociological focus* 40, 4 (2007), 436–454.

[50] Helen Nissenbaum. 2001. How computer systems embody values. *Computer* 34, 3 (2001), 120–119.

[51] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism.* nyu Press.

[52] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[53] Council of the European Union. 2008. Council Framework Decision 2008/913/JHA of 28 November 2008 on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law. (2008).

[54] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231* (2018).

[55] Caroline Criado Perez. 2019. *Invisible Women: Exposing data bias in a world designed for men.* Random House.

[56] Rashida Richardson, Jason Schultz, and Kate Crawford. 2019. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online, Forthcoming* (2019).

[57] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 59–68.

[58] Eugenia Siapera, Elena Moreo, and Jiang Zhou. 2018. *Hate Track: Tracking and Monitoring Online Racist Speech.* Technical Report. Irish Human Rights and Equality Commission.

[59] Margaret A Simons. 2001. *Beauvoir and The Second Sex: Feminism, race, and the origins of existentialism.* Rowman & Littlefield Publishers.

[60] Jesper Simonsen and Toni Robertson. 2012. *Routledge international handbook of participatory design.* Routledge.

[61] John Solomos. 1991. Les Back (1996) Racism and Society. *Hampshire, Macmillan* (1991).

[62] Ryan Steed and Aylin Caliskan. 2020. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. *arXiv preprint arXiv:2010.15052* (2020).

[63] David Sumpter. 2018. *Outnumbered: From Facebook and Google to Fake News and Filter-bubbles – The Algorithms That Control Our Lives.*

[64] Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*. 13209–13220.

[65] Robert Thorstad and Phillip Wolff. 2019. Predicting future mental illness from social media: A big-data approach. *Behavior research methods* 51, 4 (2019), 1586–1600.

[66] Gavan Titley. 2019. *Racism and media.* SAGE Publications Limited.

[67] Shannon Vallor. 2016. *Technology and the virtues: A philosophical guide to a future worth wanting.* Oxford University Press.

[68] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088* (2019).

[69] Françoise Vergès, Ashley J. Bohrer, and the author. 2021. *Front Matter.* Pluto Press, i–iv.

[70] Sandra Wachter. 2020. Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Journal* 35, 2 (2020).

[71] Marilyn Waring and Gloria Steinem. 1988. *If women counted: A new feminist economics.* Harper & Row San Francisco.

[72] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems: Gender, race and power in AI. *AI Now Institute* (2019), 1–33.

[73] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI now report 2018.* AI Now Institute at New York University New York.

[74] Langdon Winner. 1980. Do artifacts have politics? *Daedalus* (1980), 121–136.

[75] Gerhard Wohlgenannt, Ekaterina Chernyak, and Dmitry Ilvovsky. 2016. Extracting social networks from literary text with word embedding tools. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. 18–25.

[76] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876* (2018).