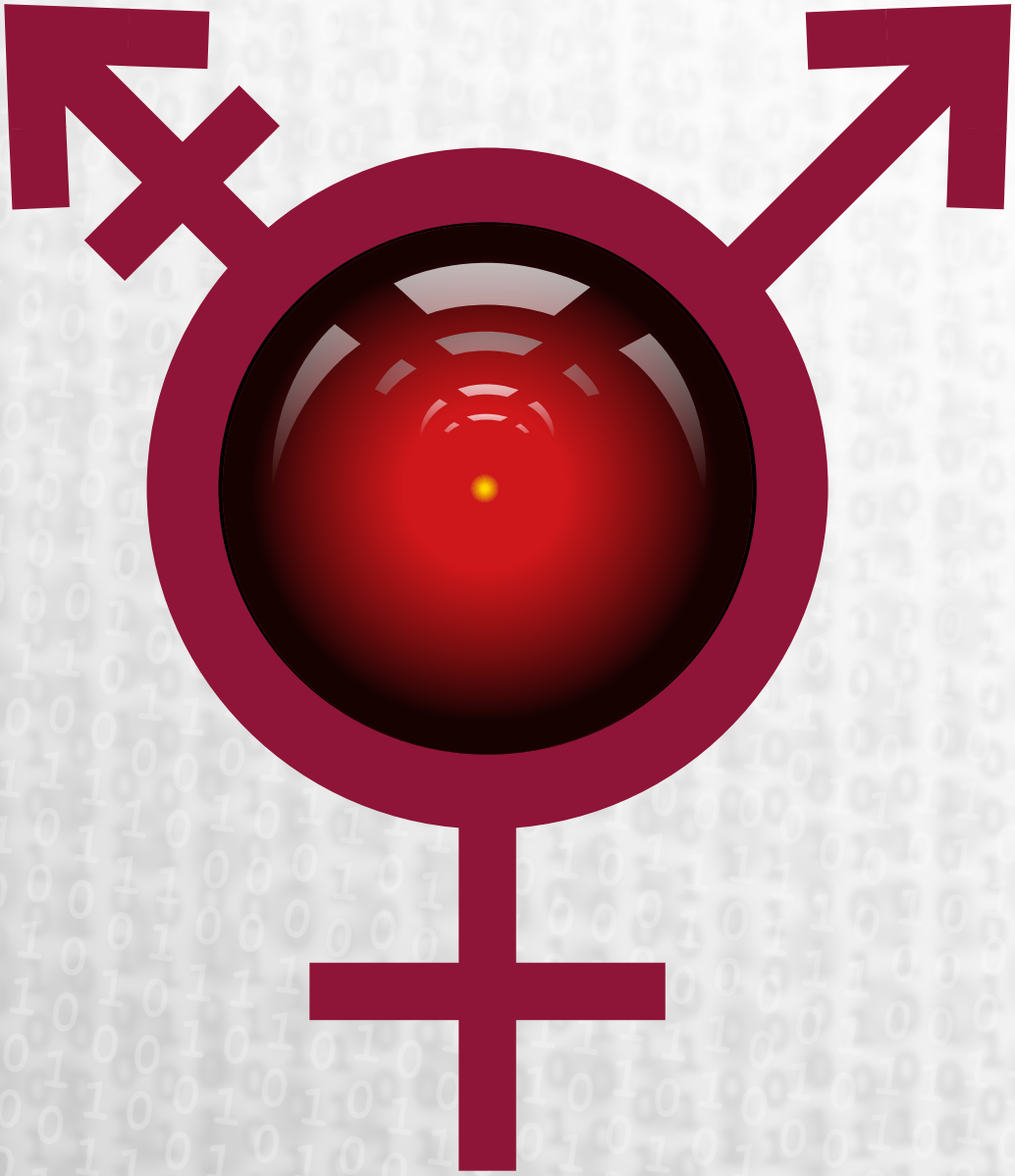


AI and Gender

Four Proposals for Future Research

Clementine Collett & Sarah Dillon

UNIVERSITY OF CAMBRIDGE

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Authors: Clementine Collett and Sarah Dillon

With administrative assistance from: Gaenor Moore

Citation: Collett, Clementine and Dillon, Sarah (2019). AI and Gender: Four Proposals for Future Research. Cambridge: The Leverhulme Centre for the Future of Intelligence.

The Reprographics Centre at the University of Cambridge designed and printed this report.

Acknowledgements

We are grateful to the Ada Lovelace Institute which provided sponsorship for the Leverhulme Centre for the Future of Intelligence 'AI and Gender' workshop. The Ada Lovelace Institute is an independent research and deliberative body – established by the Nuffield Foundation – with a mission to ensure data and AI work for people and society. The views expressed in this report are those of the authors and not necessarily of the Ada Lovelace Institute or the Nuffield Foundation.

We are grateful to PwC, who also generously supported the Leverhulme Centre for Future of Intelligence 'AI and Gender' workshop.

The Leverhulme Centre for the Future of Intelligence is grateful to the Leverhulme Trust for their generous support.

Research Centre Grant No. RC-2015-067

Thank you to all the speakers and participants at the workshop for their contribution and discussion.

Thank you to the peer reviewers for their constructive comments, we are extremely appreciative of their time and insights. Reviewers: Gina Neff, Reema Patel, Jude Browne, Lauren Wilcox, Stephen Cave, Joy Rankin, Os Keyes, Diana Robinson and Mustafa Ali.

The authors would also like to give special thanks to the following people for their contributions to the research, writing and/or revision of the report: Jude Browne, Os Keyes, Julian Huppert, the 2017-18 MPhil in Gender Studies Cohort at the University of Cambridge.

Contents

Executive Summary

This report outlines four of the weightiest challenges to gender equality presented by recent developments in artificial intelligence (AI). In tandem, it outlines four research proposals which would effectively tackle these issues.

These four research proposals are the direct result of the 'AI and Gender' workshop held by the Leverhulme Centre for the Future of Intelligence at the University of Cambridge on Wednesday 20th February 2019. This was convened with the Ada Lovelace Institute, and supported by PwC. In particular, this report draws on the final activity of the day, the collective intelligence activity. During this activity, participants (Appendix 2) were invited to identify areas and methods for future research. The report develops and augments the ideas shared during this exercise by drawing on content from the workshop presentations, questions, and discussions, as well as from a broad range of wider literature and research.

As much as this report aims to be informative, it is not intended to be prescriptive. Rather, the hope is that it will provoke action to address issues of injustice. Although this report primarily focuses on gender, rather than race, ethnicity or sexuality, it recognises the inseparability of these topics. The report advocates that future research should aim to be highly collaborative with other work in the field, and should strive to be intersectional, pluralistic, interdisciplinary and trans-sectoral. In addition, although this work situates examples mainly in the UK context, we advocate that research should be as international as possible.

Each section begins with a research context. This outlines a particular set of issues which need to be addressed and scopes the current landscape of work which is already being done to address these issues. This is followed by a summary of the proposed research, alongside indicative research methods, aims, and the value/challenges of the proposed research.

1. Bridging Gender Theory and AI Practice

Technological design often captures and reproduces controlling and restrictive conceptions of gender and race which are then repetitively reinforced. The parallel between the insistence of AI to repeat particular actions, and gender's root in repetitive social performance, mutually reinforces the restrictive mechanisms of the gender binary and racial hierarchies. We explore three notable AI systems, or aspects of systems, which repetitively reproduce controlling and restrictive conceptions of gender and race: humanoid robotics; virtual personal assistants (VPAs); and, gendered epistemology.

In order to address these issues, the report proposes research which utilises gender theory, including trans, non-binary, queer, feminist, as well as postcolonial theory, to explore the fundamental barriers to equality embedded in the design and purpose of AI technologies. In addition, this research would include assessment of areas where AI technology should not be used. This research would also pursue multilateral conversations with international stakeholders, technologists and designers, seeking to understand the conceptions and definitions of gender and race embedded in technological design.

2. Law and Policy

Laws and policies surrounding AI are currently at the embryonic stage of development. There is a risk that economic prosperity and political power will play an underlying role in shaping laws and policies concerning AI, at the expense of other more socially equalising motivations. There has been an abundance of work on how ethical codes should inform our technological practice by holding human values at the heart of development. However, there has been little work on how these can be translated into practice and embedded in policy and legislation. It goes without saying that these structures will play a crucial role in how AI shapes our world.

The report proposes that there is a need for research which analyses existing and emerging legislation and policy related to AI which will have an impact on gender equality. Specifically, this research could include policies surrounding data and privacy, technological design, and labour. These areas in particular will impact gender equality.

3. Biased Datasets

Datasets are often un-representative of the public demographic. There is a high level of data deprivation when it comes to capturing vulnerable groups. Biased datasets amplify gender and racial inequality and project past and present biases into the future.

The collection, handling and purpose of large datasets need to be further explored and exposed with regard to how these processes are perpetuating gender and racial bias and discrimination. Context-specific, gender-specific guidelines for best practice regarding data need to be established. Guidelines would cover data collection, data handling and subject-specific trade-offs. The report suggests three contexts of data use which will most starkly and significantly impact issues surrounding gender equality. These include crime and policing technologies, health technologies and financial sector technologies. The underlying social narratives of the biases present in datasets also need to be pinpointed and tackled through further research.

4. Diversity in the AI workforce

Currently, there is significant gender disparity in the AI workforce. Those designing, coding, engineering and programming AI technologies do not exhibit a diverse demographic. Nor does the current pipeline promise a better balance in the future. Gender and ethnic minorities are still not balanced in STEM subjects at school or university. Diversification of the AI workforce will be vital in order to design and implement technology which is equitable. This becomes even more urgent as there is increased demand for skilled technological experts accompanying the rise of AI. At the current rate, existing inequalities will only be aggravated and enlarged by an AI labour market which fails to reflect a diverse population.

There is a need for research which explores the factors that impact diversity in STEM education and in the AI workforce. In addition, when it comes to eliminating bias, there tends to be a reliance upon balancing numbers. Although the numbers are certainly important, research should also consider how to create a sustainable culture of diversity which can be embedded in educational institutions and in the workplace.

At present, AI technologies are repeating, perpetuating and introducing gender-based discrimination. These four proposals outline research that would address the most significant challenges which AI currently poses to gender equality. They are intended to inform and provoke practical action to improve the impact of AI on gender equality.

Introduction

Purpose and outline of the report

This report consists of four areas of proposed academic research. This research would tackle some of the most compelling challenges that recent development of artificial intelligence (AI) pose for issues pertaining to gender equality. The proposals are not intended to be prescriptive, but rather, provocative. The report aspires to use these proposals as a mechanism to raise awareness, summarise the current challenges, and prompt practical action.

The four research proposals are the direct result of the 'AI and Gender' workshop held by The Leverhulme Centre for the Future of Intelligence at the University of Cambridge on Wednesday 20th February 2019. This was convened with the Ada Lovelace Institute, and supported by PwC. In particular, the report draws on the final activity of the day, the collective intelligence activity. During this activity, participants (Appendix 2) were invited to identify areas and methods for future research. The report develops and augments the ideas shared during this exercise by drawing on content from the workshop presentations, questions and discussions, as well as a broad range of wider literature and research. All points made by individuals during the workshop are referenced as follows: (Surname, LCFI-AIG, 2019).

Each section begins with a research context. This outlines a particular set of issues which need to be addressed and scopes the current landscape of work which is already being done to address these issues. The content is largely drawn from the knowledge gained at the conference. Therefore, it is an indicative, rather than an exhaustive, account of current work in the field. Following this, there is a summary of the proposed research, alongside indicative research methods, aims, and the value/challenges of the proposed research.

Throughout, the report acknowledges scholars, organisations and institutions which are already effectively tackling particular issues. The ambition is that these research proposals will be viewed as a tool which incorporates and complements existing work, while highlighting the areas which still require investigation. Although these research proposals are contained in four separate sections, they are not mutually exclusive. Much of this work could run in parallel or could be effectively combined.

The report primarily focuses on gender, rather than race, ethnicity or sexuality, but throughout there is recognition that the issues surrounding each of these very often converge. Crucially, research should aim to attend to the intersectional, pluralistic and interdisciplinary. Although situated primarily in the UK context, mainly in order to ground the proposals in specific examples, this report advocates that research should be as international as possible.

As we continue to see rapid development of AI systems, now is the moment to address the challenges which AI presents to gender equality. The report consolidates the aims of the 'AI and Gender' workshop. It scopes and situates current research and interventions, identifies where further research and intervention is required, and acts as a call to action to tackle issues of injustice.

The AI and Gender workshop

The AI and Gender workshop was trans-disciplinary and trans-sectoral. It gathered together scholars from a wide range of fields, including computer science, history, philosophy of science, law, politics, sociology, philosophy, literature, and gender studies. In addition, it brought together researchers and practitioners from industry and research centres outside of academia, as well as key figures from UK AI governance and policy.

The aims of the conference were:

- To scope current research and interventions in the field of AI and gender;
- To situate current research and interventions in relation to wider fields;
- To collectively identify where further research and intervention is required through a collective research agenda

The day consisted of seventeen ten-minute talks from each of the speakers (Appendix 3). These took place over four panels:

- History, Narrative, Theory: Interdisciplinary Perspectives
- Trust, Transparency and Regulation
- Organisational Initiatives to Increase Gender Equality
- Challenging Built in Bias and Gender Stereotypes

The speakers shared cutting-edge research on the relationship between AI and gender. Each panel was followed by thirty minutes of questions. The day culminated in the collective intelligence activity. This addressed the third aim of the conference: to collectively identify where further

research and intervention is required through a collective research agenda.

Attendees were divided into four groups (Appendix 1) according to their expertise and a broad literature review on the work currently being done concerning AI and gender. These groups covered four topics:

- The Gender Binary: Epistemological, Physiological and Linguistic Gender Stereotypes in AI
- The Gender Politics of AI: Ethics, Policy and Privacy
- Data, Discrimination and Diversity
- AI and Gender in Organisations

The groups were asked to design at least three recommendations for new areas of research concerning AI and gender. For each of the recommendations, the groups considered:

- New directions for research/where the gaps in research are;
- Ideas on how to approach and carry out this research;
- Why this research is important.

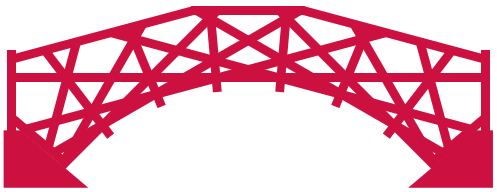
The report develops and augments four ideas shared in these four groups by drawing on content from the workshop presentations, questions and discussions, as well as from a broad range of wider literature and research.

Defining key terms

Gender:
Gender refers to the historically inherited, socially constructed, and normalised behaviours, characteristics and appearances which operate to define people as female or male, or which act as a framework to be resisted. Whilst gender can align with biological distinctions and differences, it does not necessarily do so. In this report, gender is understood to have an inextricable relationship with unequal power dynamics, and to function intersectionally with other protected characteristics such as race, ethnicity, and sexuality. When the report refers to feminist work or women's rights, this should be interpreted as mutually inclusive of trans, queer, and non-binary equality.

Artificial intelligence (AI):
Artificial intelligence is used in this report to refer to a heterogeneous network of technologies - including machine learning, natural language processing, expert systems, deep learning, computer vision, robotics – which share in common the automation of functions of the human brain.[1]

[1] This definition of AI is informed by Marcus Tomalin's introductory talk at the workshop, 'The Future of Artificial Intelligence: Language, Gender, Technology', 17 May 2019, University of Cambridge.



Research Theme 1:

Bridging Gender Theory and AI Practice

Research context

AI has a significant and profound impact on the way that people are perceived and treated in society. Yet, the design and implementation of AI perpetuates a vicious cycle. The technology captures and reproduces controlling and restrictive conceptions of gender and race which are then repetitively reinforced:

> *Gender relations can be thought of as materialised in technology, and masculinity and femininity in turn acquire their meaning and character through their enrolment and embeddedness in working machines. (Wajcman, 2010: 149)*

This mimics the repetitively reinforced nature of performative gender which we see so prominently in the work of Judith Butler. Butler's work theorizes how gender is constituted in temporal repetition; it is an action which requires a repeated performance and "[t]his repetition is at once a re-enactment and re-experiencing of meanings already socially established" (Butler, 1990: 191). Not only do we experience this through the nature of AI, which functions by repeating the same process over and over again, but also this repetitiveness is amplified by the increasing magnitude of AI development across the globe.

Feminist scholars of science and technology have been looking at the mutual shaping of gender and technology for several decades (see Shapiro, 2010 and Wajcman, 2007). Halberstam (1991) recognises that Alan Turing's 1950 paper, 'Computing Machinery and Intelligence', argued that a computer works according to the principles of imitation, but is also able to learn new things. Halberstam draws the parallel that gender is also "learned, imitative behavior that can be processed so well that it comes to look natural" (Halberstam, 1991: 443).

Lauren Wilcox has recently expanded on this relationship between gender and technology further, providing a more intersectional approach. Wilcox recognises that AI, the gender binary and colonialism all aim to essentialise, control, fix and create a hierarchy of identity. Wilcox articulates that gender itself is part of the production of racial distinctions; it is a "racializing apparatus". Both gender and race fixate on socio-political relations in order to reproduce power structures and seek to control bodies (Wilcox, LCFI-AIG, 2019; also see Wilcox, 2017).

The work of Os Keyes also demonstrates the way in which AI technology acts to control identity. Keyes uses the example of automated facial recognition, which purports to identify someone's gender by analysing photographs of them. These technologies function by a gender recognition algorithm and a race recognition algorithm. Through its reliance on fixed notions of gender and race as systems, it is inevitably discriminatory (Keyes, LCFI-AIG, 2019, also see Browne, 2015). Buolamwini and Gebru (2018) demonstrate this in their 'Gender Shades' study on facial recognition systems. The study found that darker-skinned females were the most misclassified group with an error rate of up to 34.7%. In contrast, lighter-skinned males had a maximum error rate of 0.8%. The concept of 'classification' and 'recognition' in general ought to be questioned as a legitimate and acceptable exercise.

Keyes also discusses the way that these technologies root gender within a physiological, binary frame, essentialising the body as a source of gender (also see Hamidi et al., 2018). As a consequence, they inevitably discriminate against trans people and others. They serve as a source of "infrastructural imperialism" (Vaidhynathan, 2011), building a single, normative Western construction of gender into wider systems. Keyes makes the important point that the

aim should be, instead, to "build AI that permits plural and contextual ways of being and knowing and living" (Keyes, LCFI-AIG, 2019).

Currently, there are three notable AI systems, or aspects of systems, which repetitively reproduce controlling and restrictive conceptions of gender and race:

1. Humanoid robotics

The body is a site which can be inscribed with physiological gender norms and stereotypes. Butler's work advocates that on the surface of the body, "acts and gestures, articulated and enacted desires create the illusion of an interior and organizing gender core" (Butler, 1990: 185-6). In other words, certain appearances and ways of using the body are normalised to be consistent with the meaning of 'male' and 'female'.

Nan Boyd identifies that bodies structured to abide by widely culturally intelligible boundaries (like the ones Butler describes) tend to matter politically more than others (Boyd, 2006). Here, we see another parallel between gender and technology. In the same way that gendered bodies are deemed as politically important, humanoid robots are also recognised as such. They signal economic prosperity and are an indication of technological expertise and development. By preserving physiological gender stereotypes in robotics, this results in an accumulative elevation of the political importance of both binary gender and AI.

Humanoid robotics abide by these gendered structures; they tend "to produce and reinforce gendered bodies and behaviors" (Hicks, 2015: 5). Traditional conceptions of the female body are repeatedly propagated through new technology and media (White, 2015). Londa Schiebinger demonstrates that appearance, voice, mannerisms, movements and demeanors which robots employ imitate gender stereotypes present in society (Schiebinger, LCFI-AIG, 2019; also see Hird and Roberts, 2011). Sex robots, for example, reproduce physical gender stereotypes, as well as actualising the 'objectification' of gendered bodies (Varley, 2018). Or consider Cortana, a character in the Halo video game series. An AI construct created from the cloned brain of a female scientist, Cortana has no physical form but is highly sexualised when projected as an embodied representation (Ní Loideáin, LCFI-AIG, 2019). Sophia, the humanoid robot developed by Hanson Robotics, also holds an incredibly lifelike resemblance to a stereotypical woman. Gendered bodies in robotics, particularly those of women, maintain and reproduce stereotypical appearances. Not only this, but as we can see from Sophia's Saudi Arabian citizenship, they are labelled as politically important.

Lauren Wilcox notes that gender fixes bodies in two ways: locating them in time and space by surveillance, and by framing bodies in the sense of 'correcting a problem' through the elimination of bodies that draw a threat to gendered order (Wilcox, LCFI-AIG, 2019). These humanoid representations reproduce stereotypes and, in doing so, eliminate bodies which defy gendered order.

2. Virtual personal assistants (VPAs)

Rachel Adams examines the way in which VPAs fail to criticise the binary categories of male and female. They facilitate these gender stereotypes through the power of language and naming. VPAs reproduce the concept of the female figure as the faithful aid of humankind. Without the ability to attain self-determined subjectivity of its own, the VPA is in existence only to support and assist. The VPA is literally called into being: 'hey Siri', 'hey Alexa'. Adams parallels this with Butler's theory of interpolation, which underlines how naming brings something into being and creates a power dynamic (Adams, LCFI-AIG, 2019).

The feminine voice of VPAs is associated with servitude and power disparity, and this gendering presents concerns with regard to societal harm (Bergen, 2016; Ní Loideáin and Adams, 2018; Dillon, forthcoming). Linking the language of assistance with a feminine voice has damaging implications. Woods (2018) analyses how these gender stereotypes harm society, enable surveillance, and further domesticate the feminine persona through promoting "digital domesticity" (Woods, 2018: 335, also see West, Kraut and Chew, 2019).

3. Gendered epistemology

The theory of 'intelligence' and the epistemology operationalised by AI research focuses only on a specific form of knowing. Feminist work has demonstrated that this excludes other epistemologies, including those traditionally gendered as female or feminine (Adam, LCFI-AIG, 2019; also see Adam 1995, 1998). Critical race theory has also analogised its equivalence to the knowledge of the white man (Ali, 2019; Mahendran, 2011). AI therefore takes part in a wider socio-technical exclusion or repression of women's knowledge and reifies a gendered and racialized conceptualisation of 'intelligence' (Davies, 2019). Work on AI and epistemology must countenance the possibility that AI and epistemic justice is an illegitimate combination.

Alternatively, AI may provide an opportunity to shift assumptions about male and female epistemology. For example, the narrative of 'hard' and 'soft' intelligence is often gendered as masculine and feminine respectively. Adrian Weller notes how this 'hard' intelligence, often thought to encompass logic and rationality, is much easier to reproduce in technological form. Subsequently, it is reinforced as encompassing all 'intelligence' by the fact it is adopted in these machines, but what has hitherto been thought of as 'soft' intelligence may become more privileged in being harder to encode (Weller, LFCI-AIG, 2019). As Sarah O'Connor's notes:

> *As machines become better at many cognitive tasks, it is likely that the skills they are relatively bad at will become more valuable. This list includes creative problem-solving, empathy, negotiation and persuasion. (O'Connor, 2019)*

She goes on to say that these qualities have historically "been more identified with – and encouraged in – women".

Whether AI is thought to depend upon and epitomize a masculinist epistemology, or whether AI promises to give a feminine epistemology the advantageous position in the job market, AI is perpetuating and reinforcing binary, gendered stereotypes of epistemology.

Proposed research

Although feminist theory has often been applied to technological practice (Adam, 1998), approaches to gender in technology have been critiqued by trans writers for their ignorance of trans lives (Keyes, 2018; Spiel, Keyes and Barlas, 2019). The use of gender theory needs to be broadened to further apply trans, non-binary, queer, and postcolonial theory to explore the fundamental barriers to equality embedded in the design and purpose of technologies. In addition, research would consider how we could replace these aspects with alternative, inclusive practices, or recommend against the use of AI in certain contexts altogether.

Achieving the goal of social equality would be aided by dialogue between gender theorists and technologists. But at present, gender theory and AI practice "are speaking completely different languages" (Leavy, LCFI-AIG, 2019). Susan Leavy points out that, currently, the people who are reading gender theory are not the co-authors of papers such as 'Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neutral Networks' (Alvi et al., 2018).

Gina Neff highlights this problem of the growing distance between those who are designing and deploying these systems, and those who are affected by these systems. What will ordinary people do to respond to challenge, adapt and give feedback that will be crucial for the positive evolution of these systems? Neff refers to the importance of the 'social shaping of AI', which would include designing workshops with users and including them in the discussion of how systems could be adapted to work for their benefit (Neff, LCFI-AIG, 2019).

In light of the problems outlined above, some aspects of technology in particular need to be challenged:

- The reproduction of stereotypical gendered and racialized bodies in robotics. Currently, there is a disjunction between the theoretical 'suspension of gender' which could be promising for destroying restrictive gender stereotypes, and the encapsulation of physical gender stereotypes in technology and robotics.
- The use of language, interaction and communication in relation to these systems. This refers to both the use of gendered 'voices' and 'responses' which these systems produce, in addition to the use of gendered pronouns and syntax. Susan Leavy points out that while some recent studies have sought to remove bias from learned algorithms, they largely ignore decades of research on how gender ideology is embedded in language. The mechanisms which reinforce this gendered language in technology include, for instance: the way in which certain genders are named, ordered, and described, as well as the frequency with which they are referred to, and the metaphors used to describe them (Leavy, LCFI-AIG, 2019; see also Leavy, 2018).
- The reinforcement of societal understandings of gendered epistemology in AI systems. It must challenge such structures and incorporate the insights of queer epistemology into technology.

Alongside theoretical analysis, research would also pursue a multilateral conversation with international stakeholders, technologists and designers. Seeking to understand the conceptions and definitions of gender and race, and why/how they are being embedded into technological design, would be crucial to assessing how theory can speak to practice. In turn, these would be combined to produce a set of research-based tools which could be employed by designers and technologists to embed pluralism and inclusion into AI systems, or to suggest against the use of specific AI systems altogether.

The Stanford University initiative, 'Gendered Innovations', directed by Londa Schiebinger, is focusing on integrating sex and gender research into technology. This inclusion of sex and gender analysis into "basic and applied research produces excellence in science, health & medicine, and engineering research, policy, and practice" (European Commission, 2013: 8; also see Schiebinger and Schraudner, 2011). These vital pursuits need to be extended and amplified in relation to AI technologies in order to make the design, marketing, advertisement and the ultimate purpose of these systems work for social justice.

Techno-utopianism has been criticised for coming from a place of white privilege (Ali, 2019). In addition, earlier feminist hopes that the online and virtual world would provide a site for freedom from gender constraints and inequality (Haraway, 1985 & 1996; Wajcman, 2006) have, with time, been unfulfilled. However, there is still potential for creative engagement with these AI technologies which might be used to challenge stereotypes surrounding gender (An[O]ther {AI} in Art, 2019; Dvorsky and Hughes, 2008; Ferrando, 2014; Shapiro, 2010).

Research aims

- To explore how queer, trans, non-binary, feminist and postcolonial theory shed light on practical mechanisms of discrimination and bias in existing and emerging AI technologies.
- To converse and connect theorists and technologists from a broad range of cultural backgrounds to consider global perspectives regarding definitions of gender and race, reflecting on how this is being embedded in technological practice.
- To translate this research into a relatable set of critical tools regarding AI systems. This would cover the overall purpose of systems, design, marketing, advertisement and use/distribution. In addition, it would explore where systems should not be deployed.

Indicative research methods

1. Theoretical analysis

Queer, trans, feminist, postcolonial, and other non-essentialist gender theory would be used to analyse the design and functionality of current and emerging AI technology.

In terms of embodiment, the relationship between transgenderism and embodiment would be explored in relation to AI (Elliot, 1998; Stanley & Smith, 2011). It would also be enlightening to consider how the female body can be constructed and perceived, and the role that power and politics plays in structuring and training bodies into certain gendered roles and behaviours (Butler 1990; Foucault, 1975; Grosz, 1994).

Focusing on intersectionality, it would be useful to concentrate on how racial ideology shapes gender narratives and discourse (Collins, 2005; Oyèwùmí, 1997; Spivak, 1988a&b) and, in addition, focus on the intersection of race, queer and/or trans identity, and how relevant these historical narratives are in contemporary society. These intersections and relationships could be used to reflect on how such narratives could be considered in AI technology (hooks, 1996; Puar, 2007; Snorton, 2017; Stone, 1992; Stryker 2004, 2008; Stryker & Whittle, 2006; Ware, 2017).

The queer concept of continual disruption of repetitively reinforced, historically meaningful gender norms could also be explored, especially in relation to how this could be reflected in AI systems (Ahmed, 2006; Barad, 2011; Butler, 1990, 1993; Campbell & Farrier, 2015; Eng, Halberstam and Munoz, 2005; Freeman, 2011; Kember, 2016; Warner, 1999).

2. Trans-sectoral communication

Conducting a multi-lateral conversation with international stakeholders, technologists and designers would encourage the researcher to understand contemporary non-western conceptions of what defines and frames gender and race and how these definitions are informing technological practice. It must assess how these fit into data, computing, design and implementation of AI systems that have, or will have, a global impact, despite the regional focus of their development.

There is significant theory available that can be mobilised to address technological challenges. Researchers literate in this theory need to engage with those who are designing and implementing AI in order to develop dialogue, understanding and progress.

3. Synthesis and recommended tools

Bridging the gap between gender theory and AI practice would require synthesising the theoretical work and the communicative work to produce research-based, practical tools. The aim would be for these tools to be employed and incorporated into the way that these technologies are designed and used in society. These tools could inform the technological process at all stages, as well as the more political aspects of technological creation. This would include data gathering; algorithm design; the purpose of technology; technological use; implementation and distribution. In addition, it would explore where systems should not be deployed or would be inappropriate given the goal of social justice.

Challenges

- There may be challenges when it comes to synthesising the theoretical work and the material gathered from conversing with technologists. This multi-source approach, however, will also introduce a valuable insight into the nature of any disjunctions between theoretical and practical definitions of gender and race in AI systems.
- There may be challenges in deciding how these tools can be framed in order to be easily implemented by technologists.

Value of research

- This research will foster international collaboration and networks, looking to create unity around an ambition for social equality and justice in relation to technology and its implementation.
- The research will gather an entirely new, unique evidence base of cultural understandings about how technology intersects with gender and race.
- These technologies are often exclusive, restrictive and controlling in relation to gender and race. This research would expose the ways in which this is happening and seek to open up pluralistic and inclusive ways that technology can be developed.

Research Theme 2:

Law and Policy

Research context

The development and rise of AI is often perceived to drive economic growth and intensify political power. In February 2019, President Trump signed an Executive Order which urged the continuation of American leadership in AI to motivate economic growth:

> *It is the policy of the United States Government to sustain and enhance the scientific, technological, and economic leadership position of the United States in AI R&D and development. (Executive Order No. 13,859, 84 Fed. Reg. 3967, 2019)*

The ambition for economic growth through technological innovation is also dominant in UK politics. Jude Browne notes that the UK government has not yet established a public body for AI resembling, for example, the Human Fertilisation & Embryology Authority (HFEA), which would bridge the gap between the public, experts and government. Browne argues that this encapsulates the supremacy of the private interest over the public interest, driven largely by goals of economic prosperity (Browne, forthcoming).

There is a risk that economic prosperity and political power will play an underlying role in shaping laws and policies on AI at the expense of other more socially equalising motivations. Martha Nussbaum argues that we are living in an "era dominated by the profit motive and by anxiety over national economic achievement", whereas ultimately, "it is people who matter". Nussbaum notes that policy has enormous power to shape social structures, but argues that so far, the theories which direct it largely abandon the institution of equal human rights, dignity and social justice (Nussbaum, 2011: 185-186).

Laws and policies surrounding AI are currently at the embryonic stages of development. In June 2019, the High-Level Expert Group on Artificial Intelligence for the EU will put forward policy and investment recommendations about how to strengthen Europe's competitiveness in AI. This will be one of the first steps towards developing solid policy on the topic. Up until this release, a number of ethical codes which aim to inform technological practice have been produced. These often aim to hold human principles, values and wellbeing at the heart of all developments (Dignum, 2018; Chadwick, 2018; The European Commission's High-Level Expert Group on Artificial Intelligence, 2018; Winfield and Jirotka, 2018). Nevertheless, there is, "little evidence that those principles have yet translated into practice" (Winfield and Jirotka, 2018: 9).

Structures formed through policy and legislation will be crucial in shaping the impact of AI on social equality and discrimination. Ní Loideáin and Adams (2018) recognise that there is a significant relationship between structure and discrimination:

> *Past experience in the field of regulating against sex discrimination has shown that equality can only be achieved by specific policies eliminating the conditions of structural discrimination. (Ní Loideáin and Adams, 2018: 23)*

Given their instrumental weight, it is imperative that laws and policies surrounding AI are approached and surveyed from the perspective of gender equality.

There are three areas of law and policy related to AI which will particularly impact the distribution of power and gender equality:

1. Data and privacy

The rise of AI has unleashed a hunt for vast amounts of personal data. Currently, large swathes of data are concentrated in a small number of companies, such as Amazon, Facebook and Apple. Access to this data allows systems to do a range of 'personalised' tasks, from tailoring adverts to people's personal interests, to predicting mortgages from credit ratings, age, gender, race and other personal characteristics. Naturally, this raises questions surrounding privacy and freedom, as well as questions about the nature of the raw data (which will be addressed further in Section 3 on 'Biased Datasets').

Current laws and policies surrounding data already protect people's rights to a certain extent. Ní Loideáin and Adams (2018) analyse GDPR in relation to AI. In the 'EU Charter of Fundamental Rights', they show that Article 8 outlines that everyone has the right to the protection of their personal data and users of it must seek their consent. Article 21 of the EU Charter also prohibits any discrimination based on sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, politics, property, birth, disability, age or sexual orientation. It therefore recognises personal data as a fundamental human right and this includes non-discrimination. Article 35 of the GDPR also provides that where a type of processing, in particular using new technologies, is likely to result in high risk to rights and freedoms of natural persons, the controller shall carry out an assessment of the impact on protection of personal data (a DPIA). Ní Loideáin and Adams note that although this is qualified, it translates the legal standards of the EU data protection law into reality.

Nonetheless, laws and policies surrounding the collection, storage and use of data need to be analysed further. Currently, data is being used in advertising, education and policing to reinforce racism and amplify inequality (Eubanks, 2018; O'Neill, 2016). Data is also being used to bolster our current ideologies. Filter bubbles surround us with information which aligns with our current views and deters us from engaging with ideas which conflict with our own (Chowdhury, LCFI-AIG, 2019). This amplifies the privileging of certain ideas regarding gender and race (Noble, 2018) and can "serve to exploit prejudice and marginalise certain groups" (The European Commission's High-Level Expert Group on Artificial Intelligence, 2018: 16). These uses of data need to be explored further, with special attention to the sufficiency of current regulation for vulnerable gender groups.

2. Technological design

Existing work does consider how ethical principles might be used to provide guidelines for designers in order to reduce ethical harm from the products (Winfield and Jirotka, 2018). Dignum (2017) terms this as value sensitive design: design of technology grounded in human values and situating moral questions early on in the design process.

But how do these ethical principles translate into policy and law on AI? Do they protect against gender-based discrimination? And are there regulations which deter designers from unethical design decisions?

Ní Loideáin and Adams (2018) recognise how design choices assimilate and reinforce particular stereotypes (as discussed in Section 1 on 'Bridging Gender Theory and AI Practice'). They recommend that the EU, US and UK should revise their policy documents to consider these social biases and discriminations which are integrated into the design. The European Commission's 'Draft Ethics Guidelines for Trustworthy AI' (2018) also advocate that the earliest design phase should incorporate the requirements for trustworthy AI:

> *Systems should be designed in a way that allows all citizens to use the products or services, regardless of their age, disability status or social status […] AI applications should hence not have a one-size-fits-all approach but be user-centric and consider the whole range of human abilities, skills and requirements. (The European Commission's High-Level Expert Group on Artificial Intelligence, 2018: 15)*

Technological design which considers gender equality and inclusive accessibility need to be further considered in policy and law.

3. Labour

The increased uptake of AI in organisations will significantly change the horizon of the working world. The decrease of labour intensive work is inevitable (Brynjolfsson and McAfee, 2011). Hawksworth and Berriman (2017) estimate that by 2030, around 30% of existing UK jobs will be at risk of automation. However, there is still uncertainty about what shape these changes will take; to an extent it is bound to be unpredictable and un-uniform (Form, 2015). It has been recognised that a number of policies will be influential for, or reversely will be impacted by, these changes in the labour market. Given the unpredictability of these changes and yet their significant influence on equality, it will be important to approach these policies from the angle of gender equality. This will not only include policies which influence education and training, but also policies surrounding retirement, healthcare, wages and tax, which will all be impacted through AI altering the face of the labour market (Executive Office of the President, 2016).

Proposed research

There is a need for research which analyses legislation and public policy related to AI which will impact on gender equality. So far, there has been little attention to interpreting these laws and policies through a gender lens, or indeed research into how these structures could be exploited to strive for gender equality.

Research could explore existing and emerging policies concerning AI and gender. Specifically, such research could include, but would by no means be restricted to, policies and laws surrounding three particular areas:

- Data and privacy
- Technological design
- Labour

As demonstrated in the research context, these areas are already being considered in relation to AI more generally speaking, but would benefit from additional gender-based research.

Laws and policies could be analysed through two mechanisms: (1) gender theory (2) a series of interviews with technologists, experts and policymakers.

Firstly, theoretical analysis could be used to consider how policy and legislation can facilitate AI to work for gender equality, and social equality more broadly speaking (Boden et al., 2018). Secondly, the interviews would function as a way to gain mutual understanding between policymakers and technologists regarding definitions of gender, and how vulnerable gender groups would be impacted by certain structural changes. Whittlestone et al. (2019) outline that knowledge of technological capabilities should inform our understanding of the ethical tensions, which will be "crucial for policymakers and regulators working on the governance of AI-based technologies" (Whittlestone et al., 2019: 49). This applies especially, for example, when it comes to understanding how technology will impact employment in order to design policies which tackle these changes. Collaboration between experts, policymakers and technologists would enable formation of policy which tackles the main issues in a thorough, accurate and realistic manner.

As deliverables, this research would formulate research-based, gender-specific recommendations regarding particular policies. The recommendations would suggest alterations to intricate details of existing and emerging law

and policy. Additionally, a set of guidelines for ongoing policy development would outline certain standards to be upheld when designing and implementing new policy and legislation, which both directly and indirectly impacts issues surrounding AI and gender equality. The practical element of these standards is of the utmost importance. They are not the same as an ethical framework which cannot be directly applied, but rather they would be specific, context-related and therefore straightforward for policymakers to implement.

Overall, there is a need for research which assesses how emerging and future policy structures (Erdélyi and Goldsmith, 2018) are failing to establish gender equality, and how they could be altered to strive for social justice. As Villani puts it, the ideal policy around AI will be inclusive:

> *An inclusive policy for AI must therefore incorporate a dual objective. First, to ensure that the development of AI technology does not cause an increase in social and economic inequality. Second to call on AI in order to reduce this. Rather than jeopardizing our individual trajectories and solidarity systems, AI must first and foremost help us to promote our fundamental rights, improve social cohesion, and strengthen solidarity. (Villani, 2018: 133)*

Research aims

- To analyse current and emerging law and policy which impacts the intersection of AI and gender.
- To outline specific recommendations for alterations to laws and policies surrounding AI and gender, as well as a set of research-based guidelines for ongoing policy development. These would rigorously promote the enhancement of social justice and gender equality.
- To collaborate with existing research projects, initiatives, policymakers, experts and technologists. This research aims to collaborate with existing efforts to research these areas of policy. It might inform work such as the Alan Turing Institute's research programme on 'Public Policy', which involves collaboration with policy makers on data-driven public services and innovation to solve policy problems and develop ethical foundations for data science and AI-policy making. It could also inform and collaborate with the World Economic Forum's new global initiative to assess the 'fourth industrial revolution' and how this relates to policy-making through agile governance. They define agile governance as 'adaptive, human-centred, inclusive and sustainable policy-making, which acknowledges that policy development is no longer limited to governments but rather is an increasingly multi-stakeholder effort' (Agile Governance, 2018: 4). This research aims to contribute to understanding what it means to be inclusive, especially in relation to questions of gender.
- To harness an intersectional approach and consider how these policies and laws are impacting and shaping gender, as well as race, ethnicity, sexuality, class, disability and so on. This will aid the pursuit of shaping structures in a way which considers not only one aspect of identity that could be detrimentally impacted by AI, but multiple.

Indicative research methods

1. Interviews

Interviews would be conducted with technologists and policymakers who are working in the relevant field. When interviewing technologists, it would be useful to understand how AI functions in discriminatory ways:

- Data. In addition to understanding how data is used in ways that are both visible and invisible to the public eye, and how this could be abused, interviews would cover how viable it would be to regulate such large amounts of data.
- Technological design. Interviews would focus on the process of design, seeking an insight into decision-making and which laws and policies influence these design decisions.
- Labour market. As well as seeking to understand how technology might function in the labour market, and the impact it might have on the nature of work, interviews would look to understand the greatest threats and opportunities regarding these technologies.

Interviews with policymakers would allow the research to understand processes, definitions, tensions and trade-offs which are being employed in current policy documents. Overall, interviews would enable the recommendations to be as specific and realistic as possible.

2. Theoretical analysis

Legal theory has been used in the past to analyse AI and to shape ethics surrounding these technologies (Asaro, 2007). Feminist legal theory, for example, has been employed to analyse technical issues such as privacy, surveillance and cyberstalking (Adam, 2005).

Feminist legal and policy theory could be employed to analyse gendered aspects of law and policy regarding AI. Mary Hawkesworth (1994) outlines how feminist scholarship seeks to reshape the dominant paradigms so that women's needs, interests and concerns can be understood and considered in the policymaking process. Canada, Norway and Sweden have all adopted gender and feminist-informed approaches to their foreign policies. Aggestam, Rosamund and Kronsell (2018) draw upon feminist IR theory and ethics of care to theorise feminist foreign policy. This use of gender theory could be replicated to shape policy and law surrounding AI.

In addition, anti-essentialist legal theories could be harnessed and used for analysis. In Feminist Legal Theory, Levit and Verchick (2006) outline how during the mid to late-1980s, a number of legal theorists complained about the essentialist nature of feminist legal theory. In 'Race and Essentialism in Feminist Legal Theory' (1990), Angela P. Harris argues that feminist legal theory relies on gender essentialism. This is the notion that a unitary, essential women's experience can be isolated and described independently of race, class, sexual orientation and other realities of experience. The result of this is:

> *[N]ot only that some voices are silenced in order to privilege others…but that the voices that are silenced turn out to be the same voices silenced by the mainstream legal voice of "we the people" – among them, the voices of black women. (Harris, 1990: 585)*

This research would draw on relevant legal or policy theory relating to race, gender, ethnicity, disability, sexuality, and so on in order to analyse existing and emerging policy from an intersectional perspective. This will help to ensure that a broad range of standpoints are considered when it comes to shaping AI governance. For example, this could include the use of critical race feminist legal theory, which looks at how traditional power relationships are maintained, as well as postmodern feminist legal theory, and scholars who apply queer or transgender theory to law and policy. Such research could also consider how narrative analysis might enhance traditional legal methodologies.

3. Analysis of legislation and policy

Harnessing this theoretical work alongside the interviews, this research would examine relevant laws and policies to examine their impact on issues of gender equality. This would especially be concerned with legislation and policy surrounding the key areas identified here: data and privacy; technological design; and labour.

Broadly speaking, this would isolate any content or wording which relates to how technology can facilitate inequality of power, discrimination or social injustice. Within this analysis, it could focus on:

- How these structures impact gender and racial minorities;
- The ways in which this legislation uses language and terminology to assume essentialist views of gender and race;
- The loopholes which could allow for potential inequality of power or discrimination;
- The subtext or sub-narratives in these pieces of legislation, including their assumptions of what is meant by gender and race;
- How these structures could be altered to better endorse social justice and equality.

In the case of the UK, some examples of laws which could be analysed are: Equality Act 2010; Data Protection Act/ GDPR; Digital Economy Act 2017; Policing and Crime Act 2017; Welfare Reform and Work Act 2016; Deregulation Act 2015; International Development (Gender Equality) Act 2014; Justice and Security Act 2013; Welfare Reform Act 2012.

Challenges

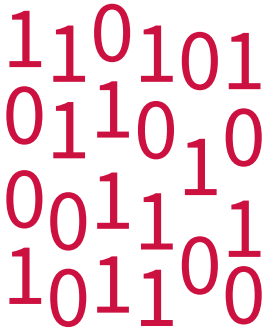
- Ensuring that technical and legal definitions of bias, equality and fairness match up with what is actually valued more broadly in society.
- Laws and policies on AI are still at the embryonic stage, which could make the process slightly staggered. However, this could also be an opportunity, especially as many policies are not yet ossified. Such research will need to keep abreast of emerging developments, and work to create access to, and inform, policies in development.
- It will be important for researcher to consider how they will address the trade-offs in terms of moral and ethical guidelines (Dignum, 2017).

Value of research

- This work will contribute to the ongoing development of policy and law surrounding AI and gender, and therefore will be influential in shaping their content and impact.
- It has been established that policy affects behavior. Structural changes implemented through this research could contribute to shifting behaviour surrounding gender equality.
- The intersectional nature of the research would enable it to consider many different standpoints, working for widespread social justice and redistribution of power.

Research Theme 3:

Biased Datasets



Research context

Mateja Jamnik recognises three sources of bias when it comes to AI systems (Jamnik, LCFI-AIG, 2019):

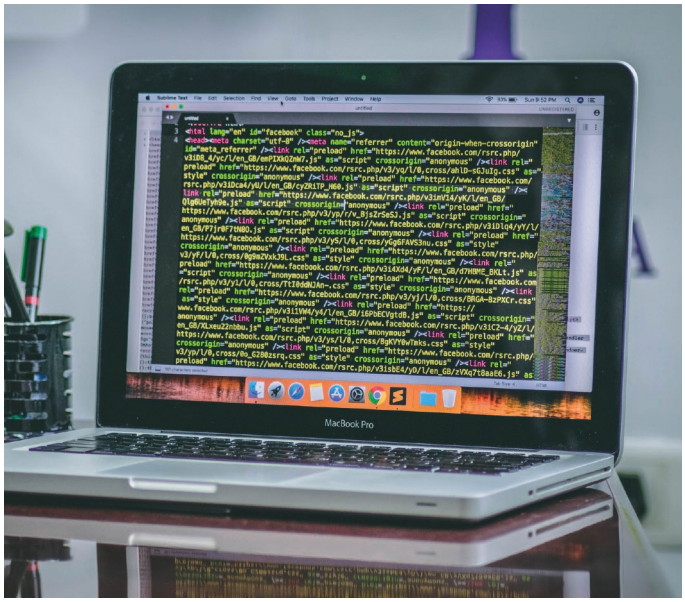
- Data. Datasets are un-representative, especially when it comes to minority groups. In some cases, this is caused by the fact that some do not have access to technology and therefore are not generating data. This means that they are not represented in the data and this propagates existing biases and exclusions.
- Algorithms. The developers, builders, engineers and installers of algorithms do not exhibit diversity. Given that it is human nature to work within one's personal view of the world, this leads to the imposition of views and values onto algorithmic systems, which in turn reinforces societal biases.
- Lack of Transparency. There is a lack of transparency; AI systems do not provide an explanation for their decisions.

Through these points, Jamnik exhibits that bias comes from people, not from machines. AI is becoming ubiquitous, so it is vital to alleviate any biases which might creep in, whether intentional or not (Jamnik, LCFI-AIG, 2019). These biases, embedded in AI systems, amplify inequality and project past and current biases into the future. This means that "AI can be a method of perpetuating bias, leading to unintended negative consequences and inequitable outcomes" (Chowdhury and Mulani, 2018).

This section will focus on the gender bias contained in datasets. The diversity of the AI workforce, a large contributor to algorithmic bias, will be addressed in greater depth in Section 4 on 'Diversity in the AI workforce'.

Datasets can take many forms and so can reproduce societal biases in multiple different ways. There are many cases from advertising, education and policing where bias data amplifies inequalities surrounding race and gender (O'Neill, 2016; Eubanks, 2018; Noble, 2018). Londa Schiebinger refers to 'The Bride Problem'. ImageNet contains 14 million labelled images, but 45.4% of these images come from the USA. This leads to biased outcomes. A white woman wearing a white wedding dress is labelled as a bride, whereas a North Indian woman wearing a wedding sari or a lehenga is labelled as performance art. These datasets simply do not incorporate or account for geodiversity (Schiebinger, LCFI-AIG, 2019, also see Zou and Schiebinger, 2018).

Datasets can also proliferate linguistic biases. Standard machine learning can acquire biases from textual data that reflect everyday human culture (Bolukbasi et al., 2016; Caliskan et al., 2017). Gendered innovations, the Stanford-based initiative, gives the example of Google Translate. When translating languages without gendered pronouns into English, the system defaults to 'he said' because the phrase



appears more on the internet than 'she said'. This reflects a faulty algorithm which selects the 'most-used' pronoun by default, but also a data-based bias (see Gendered Innovations Case Study and Bano, 2018).

These examples exhibit what Caroline Criado Perez (2019) refers to as the 'gender data gap'. Whiteness and maleness not only dominate our datasets, but they also cause a skew in datasets: "male data makes up the majority of what we know" and so "what is male comes to be seen as universal" (Criado Perez, 2019: 24). As a result, gender, racial and ethnic minorities become forgettable, ignorable and invisible. Of course, this is not the whole story. Datasets can also be disproportionately targeted at minority groups, dependent on their purpose (Gandy, 1993).

Datasets can be biased for a number of reasons (Huppert, LCFI-AIG, 2019):

- Discriminatory bias. Data has been offered, collected or handled in a subjective or inaccurate way due to discriminatory practices or personal biases. This could include the fact that data relies on people 'registering' as digital signals, and this means that certain groups are excluded from datasets. Kate Crawford summarises this: "Big data continues to present blind spots and problems of representativeness, precisely because it cannot account for those who participate in the social world in ways that do not register as digital signals" (Crawford, 2014: 1667). This discrimination results in unrepresentative datasets.
- Genuine differences. The data might be accurate and reflect genuine differences, for example, the fact that on average men are taller than women. Of course, there are also some biases which masquerade as 'genuine differences' but which are in fact rooted in societal bias (Epstein, 2007).

In many cases, then, data is "not objective, it is reflective of pre-existing social and cultural biases" (Chowdhury, LCFI-AIG, 2019). It is not enough to say that we need more representation in datasets. There is also a need to fundamentally question the raw data and what it reflects about society (Chowdhury, LCFI-AIG, 2019), as well as a need to design systems that can manage and fix biased data. This is being researched at Stanford University's Centre of Human-Centred Artificial Intelligence who are looking at how AI systems can discover and correct their own biases. In addition, we need to reflect on whether 'raw data' is an oxymoron, in that it is inevitable that it will likely be embedded within too many prior assumptions (Gitelman, 2013).

In terms of addressing biased data, Accenture have recently launched their new artificial intelligence testing service which helps companies to validate the safety, reliability and transparency of the data and algorithms in their AI systems (see Chowdhury and Mulani, 2018). The Centre for Data Ethics and Innovation, launched in 2018, is also aiming to produce best practice guidance, as well as reports with clear recommendations. These set out to build trust and enable innovation and ethical use of data-driven technologies. The Centre agrees that the ethical dimension of data and algorithms "cannot be disentangled from the context in which they are being made" (The Centre for Data Ethics and Innovation, 2019) and have chosen four sectors to explore: financial services, crime and justice, recruitment and local government. The Centre is engaged in important work on exploring algorithms and data in these sectors. However, there is still a gap in exploring these sectors specifically from the angle of gender equality.

Proposed research

The collection, handling and purpose of large datasets needs to be further explored and exposed in relation to gender. Paralleling this, the underlying social narratives of bias in these datasets need to be pinpointed and addressed.

Ethical guidelines related to AI are often non-context specific and premised on a one-size-fits-all approach (Zook et al., 2017). This research would focus on establishing more relevant, context-specific, gender-specific guidelines for best data practice. This would include contexts which relate to crime and policing technologies, health technologies, and financial sector technology. All of these examples need greater attention from the perspective of gender equality. The guidelines would cover data collection, data handling and would offer guidance on any subject-specific trade-offs.

Prior to setting these guidelines, it would be important to clarify terminology surrounding bias and fairness. This would be specifically in relation to gender, considering historical and current gender issues and tensions, which would help to form guidance regarding any trade-offs (such as context-specific instances of accuracy vs. fairness). Whittaker, Crawford, Dobbe et al. (2018) flag that work has been done to design mathematical models which are considered 'fair' when machines calculate outcomes and are aimed at avoiding discrimination. However, without a framework that

accounts for social and political contexts and histories these mathematical formulas for fairness will almost inevitably miss key factors and conceal problems which might increase or ignore issues of justice. In this case:

> *Broadening perspectives and expanding research into AI fairness and bias beyond the merely mathematical is critical to ensuring we are capable of addressing the core issues and moving the focus from parity to justice. (Whittaker, Crawford, Dobbe et al., 2018: 8)*

In other words, definitions of fairness would benefit from considerations of current and historical gender discrimination.

In parallel to this research, there is a need for research which analyses the underlying societal issues in relation to these data biases. This would look to identify the root causes of these issues: why certain pockets of society are not being captured in datasets, or why particular industries are collecting and handling data in a discriminatory way.

As previously mentioned, gender issues surrounding the use of data can differ between types of AI systems. Therefore, it would be appropriate to deal with forming data guidelines for individual contexts, rather than creating a set of 'one-size-fits-all' guidelines. These might include, but are not limited to, the following three areas, where datasets can result in significant biases, and where more work needs to be done to enhance gender equality:

1. Crime and policing technology

Predictive policing is often based on misconceptions of what variables are related to outcomes. Individuals have crimes attributed to them, as if criminality is something which is inherent to their identity (Chowdhury, LCFI-AIG, 2019). The debate in this respect focuses on classification, in particular, whether to classify and include protected characteristics such as gender and race in data which trains and informs AI systems (Whittaker et al., 2018). Corbett-Davies and Goel (2017), for example, advocate for the importance of including gender in pretrial risk assessments. Women tend to reoffend less often than men in any jurisdiction and so gender-neutral risk assessments tend to overstate the recidivism risk of women.

2. Financial services technology

AI is increasingly being used in financial services. It is being used to enhance trading systems, for example on Wall Street, where trading software makes predictions on stocks at a much faster pace than humans (Dataquest, 2017). It is also functioning to screen people for home loans, credit card loans, and to generate credit ratings.

The ethical concern is that these systems do not eliminate the bias created by humans and therefore do not exhibit 'fair' distribution of wealth or opportunity (Gokul, 2018). This could have significant implications, especially considering factors such as the gender pay gap and problematic patterns of racial wealth disparity. In these cases, algorithms which we class as 'sensible' or 'fair' which use 'representative data' may produce unsatisfactory and unjust outcomes (Huppert, LCFI-AIG, 2019).

3. Health technology

Apple's design team left out women's menstrual cycles on their health app, and the IBM Watson supercomputer was giving unsafe recommendations for treating cancer patients due to being fed hypothetical scenarios and fake patient data (Neff, LCFI-AIG, 2019; also see Chen, 2018).

Data which is fed into health technology can cause dangerous discrimination. Using a dataset of 129,450 clinical images, consisting of 2,032 different diseases, a deep convolutional neural network (CNN) can achieve performance on-par with a board-certified dermatologist. However, this dataset contains majority Europeans and Americans and fewer of those with darker skin (Esteva et al., 2017). In 2016, an analysis of genome-wide association studies showed that 80% of participants are of European

descent, a huge fail on the diversity front. However, as MIT have shown through their breast cancer screening technology, these technologies can be equitable through being accurate for racial minorities (Conner-Simons and Gordon, 2019).

Some populations are still being left behind on the road to precision medicine and this creates a privileged few in terms of access to the best medical care (Popejoy and Fullerton, 2016). Recently, DeepMind has been working on deep learning which looks for diagnosis and referral in retinal disease (De Fauw et al., 2018). They have demonstrated that performance in making a referral recommendation that reaches or exceeds that of experts on a range of sight-threatening retinal diseases after training on only 14,884 scans. Consistently, these datasets are lacking in ethnic diversity, resulting in the discriminatory outcome that most tests are significantly less accurate for ethnic minorities. This is not just an issue of social equality, but also could be a concern of life or death.

Historically, this lack of diversity in datasets used to inform healthcare has been a common trend. This has impacted not only racial and ethnic minorities, but also has greatly affected women.

Joy Rankin points out that in 1963, Maryann Bitzer came up with a way for people to learn for their nursing courses at the University of Illinois. Their lessons took place on a terminal with a television screen with a set of keys, a networked computer system known as PLATO. They learnt how to treat virtual patients with heart attacks. Rankin notes that the film they were watching was of a male patient. This highlights the issues of datasets. Heart attack symptoms are different for women and men, and, for a long-time, doctors did not know how to diagnose heart attacks in women. Rankin expands that AI could worsen health disparities. It needs to embrace intersectional data to battle structural inequality (Rankin, LCFI-AIG, 2019).

Research aims

- To align the definition of fairness (from the perspective of gender-equality) with the technical definition of fairness. Through doing this, research would aim to build a definition that can be used in analysing and addressing biased data from the perspective of gender equality.
- To produce evidence-based, context-specific, gender-specific guidelines for datasets which would help to reduce bias as far as possible and encourage fairer use of data. This would also identify where, and when, the use of such datasets should be restrained or banned.
- To pinpoint and address the societal biases underlying the biases in the datasets.
- To use an interdisciplinary approach and combine the knowledge and expertise of those collecting and handling data, companies manipulating this data and academic researchers (Shams et al., 2018). The IBM research end-to-end machine learning pipeline also demonstrates this. It recognises the fact that often the data creator, the feature engineer, the algorithm author and the user are different people. This makes the task of ensuring fairness in an end-to-end machine learning pipeline challenging (Shaikh et al., 2017). This research aims to not only encourage collaboration between theoretical and technical, but also to make the terminology and guidelines mutually accessible.
- To harness more equitable technology and ensure that the benefits of technology do not fall unfairly on particular subgroups of society. They must be shared equitably between citizens and businesses across countries and across the globe (Weller, LCFI-AIG, 2019) and delimited in their reach and deployment.

Indicative research method

1. Defining fairness and bias

Currently, there are more than twenty different definitions of fairness circulating in academic work, some focusing on group fairness and others focusing on individual fairness (Adel, LCFI-AIG, 2019, also see Kusner, Loftus, Russell and Silva, 2017). Certain definitions concern themselves with the social perception of fairness (Grgić-Hlača et al., 2018) or privacy of sensitive attributes (Kilbertus et al., 2018). Other definitions may focus on emphasising that de-biasing may leave out some of the data that is important and useful for training models, and might make the data significantly different from what it represents in the real-world:

> *As de-biasing may lead to poorly trained AI systems, at present there seems to be a situational trade-off between making sure that AI is not sexist and at the same time giving data sets that allow it to be effective in use. (Mishra and Srikumar, 2017: 69)*

Technical definitions of fairness focus on either distributive fairness (e.g. fairness of the outcome) or procedural fairness (e.g. the fairness of the decision-making process). Regarding distributive fairness, some interpretations of fairness concentrate on identifying subpopulation accuracy (Kim, Ghorbani and Zou, 2018) or aim for anti-discrimination against particular gender or racial groups (Kusner et al., 2017; Louizos et al., 2016). This might be particularly in relation to improving classification accuracies in training networks for image classifications which exhibit bias datasets (Alvi et al., 2018). Procedural fairness models include analysis of the moral judgements of input features (Grgić-Hlača, Zafar et al., 2018). This has prompted research to detect and avoid unfairness in decision-making (Dwork et al., 2012; Feldman et al., 2015; Zafar et al., 2017b; Zemel et al., 2013).

Defining fairness so that we can "transform the process into something the machine can understand" will be vital (Adel, LCFI-AIG, 2019). In this sense, more work needs to be done to relate gender equality to the definition of fairness in technology.

Considering the term 'fairness' and 'bias' in the context of gender equality, these terms could be defined through a number of methods:

- Case studies. Considering an evidence base of relevant case studies concerning datasets, these would be used to analyse how bias is reinforced and how fairness can be enhanced in various contexts.
- Gender theory. Using gender theory to inform and investigate approaches to defining bias, fairness and equality, and clarify their relationship to power and discrimination.
- Fairness tool audit. Auditing existing fairness tools to assess core ways to define fairness and bias, reflecting on how these tools might mean decisions are made that are un-biased in a technical sense, or fair in an actual sense, but still discriminatory.

2. Data and algorithms guidelines

Considering these definitions, the next step would be to formulate a set of research-informed, gender-specific guidelines. These guidelines would encompass these definitions of 'fairness' and 'bias'.

In order to make these as specific and relevant as possible to AI systems, these guidelines would be context-specific, addressing crime and policing, health, and financial services, as well as other sectors which evidently have problems with gender equality when it comes to their datasets. The guidelines would be tailored to specific uses of data in each context. These guidelines would cover methods for fair data collection; fair and transparent data handling; what a representative dataset would include; and key principles to be used when approaching trade-offs.

3. Examining societal bias

Research would analyse what data biases reveal about societal biases, specifically regarding how these could be addressed. Part of this would involve uncovering which biases are masquerading as 'genuine differences'. Analysis needs to investigate and identify the primary causes of the biases which are prevalent in society. This would consider how these biases might be alleviated, not only through data regulation and guidelines, but also through trans-sectoral and interdisciplinary societal action.

4. Audit of existing government and company initiatives

There are a number of government and company initiatives which tackle the damaging use of data in society. It would be productive to gather data on these initiatives and policies, in order to analyse their processes and impact in-depth. It would be important to explore how the frameworks of these initiatives are matching up with the results of this gender-specific research, including definitions of fairness and bias, and the context-specific guidelines. Methods for auditing initiatives will be discussed further in Section 4 'Diversity in the AI workforce'.

Challenges

- In order to carry out this research, good access to datasets would be required. This might be met with sturdy barriers given how rigorously some companies protect their data.
- There might be challenges when it comes to encouraging data scientists, large companies and academics to collaborate effectively. In many cases, these areas entail different discourses and approaches.

Value of research

- This research would establish gender-specific alignment on what defines 'fairness' and 'bias', which also looks at the entanglement and intersection of gender with race, class, ethnicity and so on. This would be of value going forward to encourage ethical data collection, handling and usage.
- Societal biases underlying dataset biases will also be pinpointed, which would aid considerations about how this discrimination could be addressed practically.
- Context-specific guidelines for data which focuses specifically on issues of gender equality, would avoid the 'one-size-fits-all' approach which generalises and does not tackle issues specific to particular AI systems.

Research Theme 4:

Diversity in the AI Workforce

Research context

"If [computer programming] doesn't sound like a woman's work – well, it just is" reads a resurfaced Cosmopolitan article from the 1960s (Burke, 2015).

In 2017, a Google employee called James Damore circulated an internal email that suggested several qualities, which he thought were more commonly found in women, including higher anxiety, explains why they were not thriving in the competitive world of coding. Google fired him, saying they could not employ someone who would argue that his female colleagues were inherently unsuited to the job (Thompson, 2019).

An overview of historical and contemporary female computer scientists is certainly sufficient evidence to bust any myths about women's contribution to the development of AI (Hicks, 2018). Ada Lovelace (1815-1852) arguably wrote the first ever code: an algorithm with which the analytic engine would calculate the Bernoulli sequence of numbers. Women were also instrumental in the development of coding and programming, especially from WWII to the 1960s. During WWII, it was principally women operating some of the first computational machines, for example at Bletchley Park (Rankin, 2015). In around 1974, a study revealed that numbers of men and women who expressed an interest in coding as a career were equal. By 1983-1984, 37.1% of all students graduating with degrees in computer and information science were women (Thompson, 2019; also see Light, 1999).

Computing/programming was initially dominated by women. Initially, it was perceived as more clerical and low-skilled work. However, as the field became more culturally, economically and socially valuable, the profession itself became perceived as more valuable. With this raised prestige, men moved in, and women were increasingly pushed out (Abbate, 2012; Barriers to Equality, 1983; Ensmenger, 2010; Hicks, 2018; Misa, 2010; Rankin, 2018).

Fast-forward over 3 decades, and only 7% of students studying computer science and 17% of those working in technology in the UK are female (Liberty – Written evidence (AIC0181), 2017: point 27). Currently, there is major gender disparity in the AI workforce. Those designing, coding, engineering and programming AI technologies do not exhibit a diverse demographic. As noted in Section 3 on 'Biased Datasets', this has a significant impact on algorithmic design, which in turn affects the biased output of AI technology. The current pipeline does not promise a better balance in the future. Gender and ethnic minorities are still not balanced in STEM subjects at school or at university.

Judy Wajcman (2010) outlines that these disparities lead to a vicious cycle: lack of childhood exposure to technology, lack of female role models, and extreme segregation in the job market all lead to women being perceived as technically incompetent (Wajcman, 2010). Gender stereotypes regarding labour, which are engrained early in

life, have significant implications, as Eagly and Wood (2012) acknowledge:

> *Such gender role beliefs, shared within a society, promote socialization practices that encourage children to gain the skills, traits, and preferences that support their society's division of labour. (Eagly and Wood, 2012: 57-58)*

Sandra Harding's work on standpoint theory recognises that in hierarchical societies, the dominant group of people produce the epistemology, social theory and the conceptual frameworks, and that these "conventional epistemologies tend to naturalize social power" (Harding, 2010: 173). In the same way, those involved in designing future technology are dictating and framing how society functions.

The characterisation of men as more suitable for jobs in STEM still seems to be prevalent. Thompson (2019) comments on the underlying current of sexism which continues to persist in technological careers:

> *The assumption that the makeup of the coding workforce reflects a pure meritocracy runs deep among many Silicon Valley men…sociobiology offers a way to explain things, particularly for the type who prefers to believe that sexism in the workplace is not a big deal, or even doubts it really exists. (Thompson, 2019)*

It is inevitable that there will be increased demand for skilled technological experts with the increased uptake of AI in society. Daugherty et al. (2018) argue that AI can help us to address biases instead of perpetuating them, but this positive effect would come from the humans who design, train and refine these systems:

> *Specifically, the people working with the technology must do a much better job of building inclusion and diversity into AI design…thinking about gender roles and diversity when developing bots and other applications that engage with the public. (Daughtery et al., 2018)*

Diversification of the AI workforce will be vital in order to design and implement technology which is equitable (Weller, LCFI-AIG, 2019; also see Hall and Pesenti, 2017), even if it this alone is not a sufficient condition for equity (Ali, 2018). Diversity brings fresh and varied perspective and encourages deliberation. Additionally, a lack of diversity exhibited by an unvarying workforce alienates those who are not consistent with this image and creates a sense of superiority for those who do fit this image. Donn Byrne's similarity-attraction theory (1969) argues that individuals are attracted to those with whom they share something in common. This psychological phenomenon has a problematic result: organisations "tend to recruit in their own image" (Singh, 2002: 3).

Sara Ahmed picks up on the occurrence that there can be "comfort in reflection" which comes from a familiarity of bodies and worlds (Ahmed, 2012: 40). It is through this comfort that workplaces extend, through reproduction, the space of the organisation by constantly replicating and reproducing the overpowering demographic through the repetitive process of recruitment. This applies to race and gender and is an important aspect of addressing diversity. When women and ethnic minorities are encouraged to apply, "the logic exercised here is one of 'welcoming', premised on a distinction between the institution as host and the potential employee as guest" (Ahmed, 2012: 42). The concept of certain genders or races being more 'at home' or 'entitled' to a workplace than others is not one which follows a narrative of social justice.

Another site of gendered normalisation has been the association of men with leadership. Hoyt and Murphy (2016) label leadership stereotypes as 'implicit leadership theories'. Broadly speaking, these lead to a 'stereotype-based lack-of-fit' between females and success within leadership positions (Heilman, 2012; Lyness & Heilman, 2006). Stereotypically male qualities are thought of as necessary to be a successful executive (Martell et al., 1998; Wille et al., 2018). These stereotypes are dangerous; they shape the way our labour market, including that of the AI industry, is structured and dominated.

At the current rate, existing inequalities will only be aggravated and enlarged by an AI labour market which fails to reflect a diverse population. There are two problematic stages of the pipeline which need to be addressed:

1. Education in STEM at school and university

Uptake of STEM subjects (science, technology, engineering, maths) at school and university still exhibit a considerable lack of diversity in most countries (Sanders, 2005).

Last year, PwC surveyed 2000 A-level students in the UK, looking at their perceptions of technology. From the sample, 78% could not name a woman working in technology. Regarding future careers, 27% of female students said that

they would consider a career in technology and only 3% said this would be their first choice. This is no surprise considering that only 6% had had it suggested to them as a career option. In comparison, 61% of male students said they would consider a career in technology.

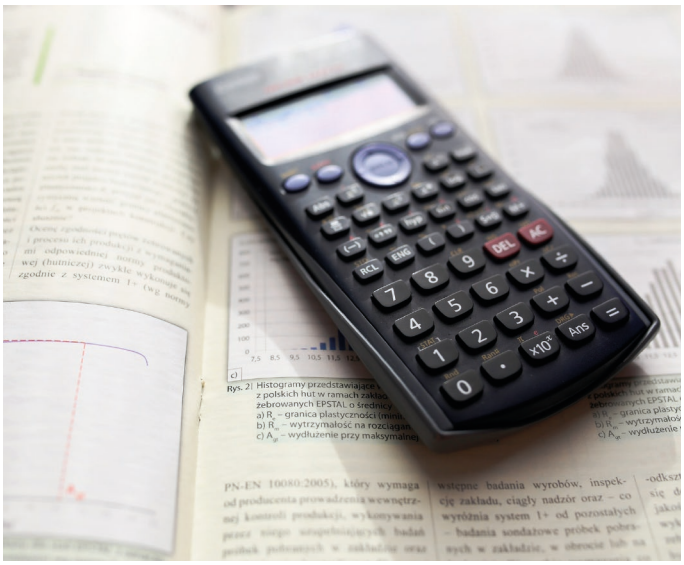
Rob McCargow from PwC recognises that these statistics highlight substantial factors contributing to the current imbalance in the technology labour market. It is clear that careers in technology have not been normalised for women. They are not commonly suggested to young women, and even though there are inspiring women working in tech, it is not well-known which roles models to look up to. If these pipeline issues cannot be addressed, things will only get worse (McCargow, LCFI-AIG, 2019).

Sinel and He (LCFI-AIG, 2019) also advocate that we need to change what we teach and how we influence young minds in the way they view technology. Many young people are not introduced to the ethical side of technology or the social implications of technology, and they do not even get to discuss and debate it (Sinel and He, LCFI-AIG, 2019).

2. Diversity of the AI workforce

In August 2018, an article was published by Element AI Lab which showed the gender ratio of AI researchers around the world. It analysed 4000 researchers who had been published in 2017 at the leading conferences on AI: NIPS, ICML and ICLR. They found that, on average, 88% of publications were written by men (Mantha and Hudson, 2018).

Career opportunities in AI will only increase as it becomes more pervasive. The Alan Turing Society has just launched a new research project, 'Women in Data Science and AI', directed by Judy Wajcman. This will inform concrete policy measures aimed at increasing the number of women in data science and AI. They predict that by 2020, more than 2.7 million data scientist job openings are forecast to be advertised in the US alone. In the UK, women represent 47% of the workforce but hold less than 17% of all technology jobs.



Proposed research

1) Investigating psychological factors surrounding diversity in STEM education and the AI labour market

There is a need for research which explores the factors that impact diversity in STEM education and in the AI workforce. In order to investigate the root causes for these inequalities, research is needed which explores the psychological elements of decision-making regarding subject choices at school and university, as well as career choices. This exploration could cover the psychological factors and biases from both sides: what factors motivate women and minority groups to pursue these subjects or occupations, and what biases impede diversity in the industry through application processes and subconscious bias in the context of recruitment:

- STEM subjects at school and university. There is a need to explore which factors play a fundamental role in encouraging or discouraging the uptake of STEM subjects for women and racial or ethnic minorities in school and universities. This would include an exploration of school curriculums, psychological motivators, university and career preparation, admissions processes and dropout rates/reasons.
- AI careers. Likewise, there is a need for more evidence on what encourages and discourages minority groups in applying for jobs in AI, as well as progressing up the promotion ladder. This would explore factors such as job advertisements, reasons for leaving the field, and promotion processes in technology companies.

2) Exploring mechanisms to embed a culture of diversity

When it comes to diversity, there tends to be a reliance upon balancing numbers in order to eliminate bias (Berenstain, 2016). Although this is certainly important, research should also consider how to create a sustainable culture of diversity through social analysis skills. These would be embedded in education and the workplace, so that it is not always about being dependent on the people in the room.

Diversity in design is not only about opportunities but about how the world is designed and for whom (Wajcman, 2007). Neither is it just a matter of increasing the technical skills of diverse groups so that they can rise to the top (Hick, 2013). As Ahmed notes:

> *Diversity would be institutionalized when it becomes part of what an institution is already doing, when it ceases to cause trouble. (Ahmed, 2012: 27)*

Sara Ahmed emphasises that diversity is not just about equalising numbers or fulfilling quotas. Diversity needs to be embedded within the organisational flow of an institution (Ahmed, 2012). In these contexts, for example, this could include thinking about which scholars could be on the reading lists, ways to title modules, or approaches to teaching.

Research aims

- To identify the main barriers to diversity in STEM subjects at school and university, and subsequently in the AI labour market. By uncovering these barriers, this research aims to discover the ways in which these might be effectively addressed by institutions and organisations. CognitionX recently asserted that, "one of the reliable ways we know we can mitigate [the problem of bias and discrimination] is to have more diverse development teams in terms of specialisms, identities and experience" (CognitionX written evidence, AIC0170, 2017: point 8.7).
- To find ways to embed cultures of diversity into institutions and organisations. Diversity is not just a number balance; it is about attitudes, behaviours and perceptions. This research should aim to examine not only what is taught but how it is taught.
- To disrupt the association of particular genders with particular lines of work (Hicks, 2013).

Indicative research methods

1. Psychological factors

A combination of qualitative and quantitative research could be used to look for any correlations between psychological factors (such as motivation and confidence), to statistical patterns which emerge. This data can be analysed:

- To discover any correlations between the qualitative and quantitative data. Causal inferences could be determined between the statistical bottlenecks and the psychological or cultural reasons for these barriers;
- To assess the ways in which diversity can become part of the infrastructure and culture of institutions and organisations;
- To explore any parallels, cross-fertilisations, and implementations of theoretical concepts of gender and race which are apparent in this material (Singler, LCFI-AIG, 2019).

Qualitative data: focus groups

Small focus groups in schools and universities would allow students to share their experience regarding subject choice, teaching and institutional culture.

These focus groups in schools would focus on questions of:

- Curriculum: what is covered in school regarding technology/AI?
- University and career choices: how are people at school introduced to future university choices and careers by teachers and mentors, and how is this gendered, if at all?
- Perception of subjects: how do students perceive STEM subjects and careers in technology to fit with different genders, if at all?
- Ability: how do different genders perceive their ability in STEM subjects?
- Motivations and discouragement: why are people motivated or reluctant to take up STEM subjects?

At university, there could be a focus on questions of:

- Motivation: are there any differences in why men and women choose to apply for STEM subjects?
- Confidence: comparatively, how confident do men and women feel about their subject?

- Culture: do people experience a culture of diversity or domination within the subject?
- Career: how does gender impact future ambitions, if at all?

Qualitative data: surveys
Surveys in universities and organisations would help to gather data regarding the university admissions process and the recruitment processes.

Within the university admission process, surveys might focus on which aspects of their process are impacting application rates or acceptances of different genders, alongside how universities are looking to tackle this and whether this is having any noticeable effect.

Regarding recruitment processes, surveys could look at how jobs are advertised and how the nature of their recruitment processes are impacting diversity. Surveys could also focus on workplace culture, reasons why people are leaving work, and promotion systems.

Qualitative data: discourse analysis
Discourse analysis methods could be used to examine rhetoric around available employment procedures and diversity statements, in conjunction with interview material from HR professionals at technology companies creating AI. This would provide a grounding in the suppositions underpinning hiring practices (Singler, LCFI-AIG, 2019).

Quantitative data collection
Research would gather data on the following, with reference to gender, race and ethnicity:

- Subject choices in schools;
- Demographic of applications for computer science and STEM subjects at university;
- Demographic of successful applicants for computer science and STEM subjects at university;
- Demographic of applicants for jobs in AI positions;
- Demographic of successful applicants for jobs in ML and AI;
- Current statistics on diversity in universities;
- Current statistics on diversity in labour market;
- Drop-out rate at university.

2. Analysis of current initiatives
Research is needed to gather an evidence base regarding the efficacy of current initiatives which aim to redress the imbalance in education and careers in STEM. This would allow for an in-depth theoretical analysis of these initiatives in terms of their structural and behavioural impacts surrounding inequality, stereotypes and opportunities.

In What works: gender equality by design (2016) Iris Bohnet advocates that randomised control trials point towards a number of evidence-based interventions which could effectively tackle problems. The interventions are subsequently tailored to people's behaviour. Bohnet outlines the importance of examining the effectiveness of these behavioural designs in the same way we would examine a drug trial: running trials in which people are randomly assigned to control groups.

Given the variation of approaches these initiatives hold for tackling issues of gender equality in the AI workforce pipeline, this methodology could be applied to analyse the impact of certain initiatives. Initiatives built on 'behavioural design' would have their effectiveness measured through randomised control trials. This would enable analysis of the impact of these initiatives in order to provide evidence of what actually work. This experimental method would also give indications of how interventions could be altered to optimise their impact. The data collected would also be used as an evidence base for 'design thinking' of alternative methods of intervention.

Sinel and He identified the most effective way to address applicant inequality by trial. Their initiative, 'Teens in AI' exists to inspire the next generation of AI researchers, entrepreneurs and leaders. They were experiencing a significant lack of diversity in applicants. Attempting to change this, they introduced gender targets and scholarships for girls. However, this did not encourage girls to join; the acceptance rate was still around 90% boys. A blind admissions process was implemented, which resulted in 50:50 selection. This led to more girls applying and more girls being admitted (Sinel and He, LCFI-AIG, 2019). Here, we see an experimental, design-related, evidence-based, positive change in their programme.

Challenges

- Due to privacy regulations, it could be difficult to gain access to information on pupils/students, or to organisational data. Even if not every school, university or organisation can be accessed, the research will be valuable as long as it has the best possible representation of the country.
- It is extremely difficult to compare recruitment processes. Each company recruits in different ways and therefore each process has its own problems and merits. However, this would be crucial for examining the problems in the workplace and how bias comes into play.
- Focus groups could be limiting in the sense that students might not feel they can be open in front of their peers. This is a challenge, but it can be addressed. For instance, researchers could try reducing group sizes or putting people in groups they are more comfortable with.
- It is worth noting that there will be variation in the nature and severity of obstacles between each school, institution and organisation. This makes it challenging to do a direct comparison. For example, in January 2018 the Department of Education recorded that there were 3.26 million pupils in state-funded secondary schools and 0.58 million in independent schools and 0.12 million in 'special schools'. These schools will exhibit significant differences (e.g. some will be single sex, some will have bigger wealth disparities) and therefore, probably, have very different issues regarding diversity. Research should consider how to tackle this, and how to formulate a study which could compare similar institutions/organisations or isolate each one to address their individual issues.

Value of research

- Research would draw connections between all stages of the pipeline, from school through to university and into the workforce. This holistic approach is unique, considering that initiatives or research often only focuses on one aspect of the pipeline.
- Identifying specific areas where statistical bottlenecks occur in schools, universities and workplaces would be valuable. The complementary qualitative research will enable causal inferences to be drawn, and a tailored solution to be explored.
- This research would have far-reaching and long-term effects in terms of disrupting current associations of gender with particular forms of labour.
- Testing initiatives through trials would highlight which methods are having a discernible impact and which techniques are not effective. This would be valuable both in terms of channeling economic investment into the most productive initiatives, and also in terms of altering less effective initiatives to increase their impact.

Conclusion

In light of the 'AI and Gender' workshop held by the Leverhulme Centre for the Future of Intelligence at the University of Cambridge on the 20th February 2019, this report has suggested four academic research proposals. These four proposals have been designed to counter some of the most urgent and significant challenges which AI currently poses to gender equality.

This report has suggested that, at the current rate, AI will continue to perpetuate gender-based discrimination. It has highlighted how this occurs through the design of AI systems which reinforce restrictive gender stereotypes; law and policy which is not focused on issues of gender equality; the widespread use of bias datasets; and a lack of diversity in the AI workforce. Future research on these issues should attend to the intersectional, collaborative and pluralistic, as well as aiming to be interdisciplinary, international and trans-sectoral.

These proposals are intended to provoke practical action on these issues surrounding the impact of AI on gender equality.

Bibliography

All online references accessed May 2019.

Abbate, J (2012). *Recoding gender: women's changing participation in computing.* Cambridge, Mass: MIT Press.

Accenture (2018) *Accenture Launches New Artificial Intelligence Testing Services.* [online] Available at: https://newsroom.accenture.com/news/accenture-launches-new-artificial-intelligence-testing-services.htm

Adam, A. (1995). Artificial intelligence and women's knowledge. *Women's Studies International Forum,* 18(4), pp.407-415.

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

Adam, A. (2005). *Gender, Ethics and Information Technology.* London: Palgrave MACM, New York NY, USA. illan.

Adel, T., Valera, I, Ghahramani, Z. and Weller, A. (2019). One-network Adversarial Fairness. In: *33rd AAAI Conference on Artificial Intelligence.* Hawaii, January 2019.

Aggestam K, Bergman-Rosamond A and Kronsell A (2018) Theorising feminist foreign policy. *International Relations* [online] 32(4), 1–17. Available at: https://journals.sagepub.com/doi/pdf/10.1177/0047117818811892 .

Ahmed, S. (2012). *On being included: Racism and diversity in institutional life.* London: Duke University Press.

Ahmed, S. (2006). *Queer Phenomenology.* Durham: Duke University Press.

Alexander, E. et al. (2014). Asking for Help from a Gendered Robot. In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society.* Quebec City, Canada, CogSci2014.

Ali, Mustafa (2019). "White Crisis" and/as "Existential Risk", or The Entangled Apocalypticism of Artificial Intelligence. Zygon: Journal of Religion and Science, 54(1), pp.207-224.

Ali, S.M. (2018) Invited Chair's Response to "AI and Epistemic Injustice: Whose Knowledge and Authority?", Session 1 of one-day workshop 'Origin Myths of Artificial Intelligence: Histories of Technology and Power', Levehulme Centre for the Future of Intelligence (CFI), University of Cambridge, 30 November 2018 (http://lcfi.ac.uk/events/origin-myths-artificial-intelligence-histories-tec/)

Alvi, M., Zisserman, A., and Nellaker, C. (2018). Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings. [online] Arxiv. [Preprint] Available at: arXiv:1809.02169

An, M. (2017). *Artificial Intelligence is here - people just don't realize it.* [Blog] HubSpot. Available at: https://blog.hubspot.com/news-trends/artificial-intelligence-is-here

An[O]ther {AI} in Art, Summit (2019). Website: https://www.anotherai.art

Antin, J., Yee, R., Cheshire, C., and Nov, O. (2011). Gender Differences in Wikipedia Editing. In: *7th Annual International Symposium on Wikis and Open Collaboration.* Mountain View, CA, US: WikiSym 2011.

Asaro, P. (2007). Robots and Responsibility from a Legal Perspective. In: *Proceedings of the IEEE International Conference on Robotics and Automation.* Rome: IEEE.

Bano, M. (2018). *Artificial intelligence is demonstrating gender bias – and it's our fault.* [online] Kings College London News Centre. Available at: https://www.kcl.ac.uk/news/news-article?id=c97f7c12-ae02-4394-8f84-31ba4d56ddf7

Barad, K. (2011). Nature's Queer Performativity. *Qui Parle: Critical Humanities and Social Sciences,* 19(2), pp.121-158.

Berenstain, N (2016). Epistemic Exploitation. *Ergo*, 3(22).

Bergen, H., (2016). 'I'd Blush if I Could': Digital Assistants, Disembodied Cyborgs and the Problem of Gender. *A Journal of Literature Studies and Linguistics,* VI, pp.95-113.

Berlant, L. (2006). Cruel Optimism. *Differences,* 17(3), pp.20-36.

Berriman, R. and Hawksworth, J. (2017). *Will robots steal our jobs? The potential impact of automation on the UK and other major economies.* [online] PwC. Available at: https://www.pwc.co.uk/economic-services/ukeo/pwcukeo-section-4-automation-march-2017-v2.pdf

Bivens, R. (2017). The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media & Society,*19(6), pp.880-898.

Bivens, R. and Haimson, O. (2016). Baking Gender into Social Media design: How platforms shape categories for users and advertisers. *Social Media and Society,* 2(4), pp. 1-12.

Boden, M. et al. (2017) Principles of robotics: regulating robots in the real world, *Connection Science,* 29(2), pp.124-129.

Bohnet, Iris (2016). What works: gender equality by design. Cambridge: Mass: The Belknap Press of Harvard University.

Bolukbasi, T. et al. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing word embeddings. [online] *Arxiv.* [Preprint] Available from: arXiv:1607.06520

Boyd, N. (2006). 'Bodies in Motion'. In: S. Stryker and S. Whittle, ed., *The Transgender Studies Reader.* New York: Routledge.

British Academy and The Royal Society (2018). The Impact of Artificial Intelligence on work. [online]. Available at: https://royalsociety.org/-/media/policy/projects/ai-and-work/evidence-synthesis-the-impact-of-AI-on-work.PDF

British Standards Institution. (2016). BS 8611:2016 *Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems.* London, UK: BSI.

Brooks, R. et al (1999). The Cog Project: Building a Humanoid Robot. In: Nehaniv C.L. (eds) *Computation for Metaphors, Analogy, and Agents. CMAA 1998. Lecture Notes in Computer Science, vol 1562.* Springer, Berlin, Heidelberg.

Brooks, R. (2002). *Flesh and machines: How robots will change us.* London: Allen Lane.

Brooks, R. (2018) *Talk – Step Towards Super Intelligence.* Department of Computer Science, University of Oxford.

Browne, J. (forthcoming, 2019). *Political Responsibility and the Public Interest.*

Browne, J. (forthcoming). '100 Years to Bliss? Artificial Intelligence, Politics and Regulation'. *Presented at the CFI Artificial Intelligence, Politics and Regulation: A Workshop. 27th September 2018.* http://lcfi.ac.uk/events/artificial-intelligence-politics-and-regulation-wo/

Browne, S (2015). *Dark matters: on the surveillance of blackness.* Durham: Duke University Press.

Brynjolfsson, E. and McAffee, A. (2011). *Race Against the Machine.* Digital Frontier Press.

Buolamwini, J. (2018). Amazon's Symptoms of FML – Failed Machine Learning – Echo the Gender Pay Gap and Policing Concerns. *Medium.* [online] Available at: https://medium.com/mit-media-lab/amazons-symptoms-of-fml-failed-machine-learning-echo-the-gender-pay-gap-and-policing-concerns-3de9553d9bd1

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency,* PMLR 81: pp. 77-91.

Burke, E. (2015). The Computer Girls: 1967 Cosmo article highlights women in technology. *Silicon Republic.* [online] Available at: https://www.siliconrepublic.com/people/women-in-technology-the-computer-girls-cosmopolitan

Butler, J. (1990). *Gender Trouble: feminism and the subversion of identity.* New York: Routledge.

Butler, J. (1993). Critically Queer GLQ: *A Journal of Lesbian and Gay Studies,* 1(1), pp.17-32.

Byrne, D. et al. (1969). Attitude Similarity-Dissimilarity and Attraction: Generality Beyond the College Sophomore. *The Journal of Social Psychology,* 79(2), pp.155-161.

Caliskan, A., Bryson, J. and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science,* 356 (6334), pp.183-186.

Campbell, A. and Farrier, S. (2015). Queer Practice as Research: A Fabulously Messy Business. *Theatre Research International,* 40(1), pp. 83-87.

Cave, S., Coughlan, K. and Dihal, K. *'Scary Robots': Examining public responses to AI.* In: Conference on AI Ethics and Society 2019. New York NY, US: AAAI/ACM.

Cave, Stephen; Craig, Claire; Dihal, Kanta; Dillon, Sarah; Montgomery, Jessica; Singler, Beth and Taylor, Lindsay (2018). Portrayals and Perceptions of AI and Why They Matter. London: The Royal Society.

Chadwick, P. (2018). To regulate AI we need new laws, not just a code of ethics. *Guardian Online,* [online] Available at: https://www.theguardian.com/commentisfree/2018/oct/28/regulate-ai-new-laws-code-of-ethics-technology-power

Chen, A. (2018). IBM's Watson gave unsafe recommendations for treating cancer. *The Verge,* [online] Available at: https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science

Chessen, M. (2018). Encoded Laws, Policies, and Virtues. *Cornell Policy Review,* [online] Available at: http://www.cornellpolicyreview.com/encoded-laws/

Chowdhury, R. and Mulani, N. (2018). Auditing Algorithms for Bias. *Harvard Business Review,* [online] Available at: https://hbr.org/2018/10/auditing-algorithms-for-bias

Coates, J. (2004). *Women, Men and Language.* London: Routledge.

CognitionX – Written evidence (AIC0170) (2017) [online] Available at: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69680.html

Collins, P. H. (2005). *Black Sexual Politics: African Americans, Gender, and the New Racism*. New York: Routledge.

Conner-Simons, Adam and Gordon, Rachel (2019). Using AI to predict breast cancer and personalize care. [online] *MIT News,* Available from: http://news.mit.edu/2019/using-ai-predict-breast-cancer-and-personalize-care-0507

Consultation outcome: Centre for Data Ethics and Innovation Consultation (2018). Department for Digital, Culture, Media & Sport. Available from: https://www.gov.uk/government/consultations/consultation-on-the-centre-for-data-ethics-and-innovation/centre-for-data-ethics-and-innovation-consultation

Corbett-Davies, S. et al. (2017). *Algorithmic decision making and the cost of fairness*. Working Paper, Stanford University.

Craig, Claire (2019). How does government listen to scientists? Switzerland: Palgrave Macmillan.

Crawford, K., Miltner, K. and Gray, M. (2014). Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication,* 8, pp.1663-1672.

D'Ignazio, C., and Klein, L. (2019) *Data Feminism* [online] Available at: https://bookbook.pubpub.org/data-feminism

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

Dataquest: Artificial Intelligence in Financial Services: Opportunities and Challenges (2017).

Datta, A., Tschantz, M. and Datta, A. (2015). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. In: *Proceedings on Privacy Enhancing Technologies,* 2015 (1) pp. 92-112.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S (2017). Proxy Discrimination in Data-Driven Systems.

Daugherty, P., Wilson, J. and Chowdhury, R. (2019). Using Artificial Intelligence to Promote Diversity. *MIT Sloan Management Review,* 60(2), pp. 10-12.

Davies, S (2019). 'Women's minds matter', [online] *Aeon,* Available at: https://aeon.co/essays/feminists-never-bought-the-idea-of-a-mind-set-free-from-its-body

Dennis, M., & Kunkel, A. (2004). Perceptions of men, women and CEOs: The effects of Gender Identity. *Social Behavior and Personality,* 32(2), pp.155-171.

Dignum, V. (2017). Responsible Artificial Intelligence: Designing AI for Human Values. *ITU Journal: ICT Discoveries,* Special Issue No.1, pp. 1-8.

Dignum, V. (2017). Responsible Autonomy. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).* [online] pp. 4698-4704. Available at: https://www.ijcai.org/proceedings/2017/0655.pdf

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. Ethics and Information Technology, 20:1-3.

Dillon, S. (in preparation). *Literature and Artificial Intelligence: Narrative Knowledge and Applied Epistemology.*

Dillon, S. (forthcoming). "The ineradicable Eliza effect and its dangers": Weizenbaum, *Pygmalion* and the implications of gendering AI'

Dillon, S. and Schaffer-Goddard, J. (forthcoming). What AI Researchers Read: The Role of Literature in Artificial Intelligence Research.

Dizikes, P. (2019). *AI, the law, and our future.* [online] MIT News: MIT Policy Congress. Available at: http://news.mit.edu/2019/first-ai-policy-congress-0118

Dobbin, F. and Kalev, A (2016). Why Diversity Programs Fail. *Harvard Business Review* 94, no. 7/8.

Dolezal, L. (2015). The Body, Gender, and Biotechnology in Winterson, J. *The Stone Gods. Literature and Medicine,* 33(1), pp. 91-112.

Dvorsky, G. and Hughes, J. (2008). *Postgenderism: Beyond the Gender Binary.* IEET Monograph Series, pp. 1-18.

Dwork, C. et al. (2012). Fairness Through Awareness. In: ITCS '12 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. Cambridge, Mass: ITCS, pp. 214-226.

Eagly. A. and Wood, W. (2012). Biosocial Construction of Sex Differences and Similarities in Behavior. *Advances in Experimental Social Psychology,* 46, pp.55-123.

Edwards, H. (2017). There's a fine line between what people want robots to do and not do for them. *Quartz.* [online] 22 October 2017. Available at: https://qz.com/1101460/theres-a-fine- line-between-what-people-want-robots-to-do-and-not- do-for-them/

Elliot, P. and Roen, K. (1998) Transgenderism and the Question of Embodiment: Promising Queer Politics? *GLQ: A Journal of Lesbian and Gay Studies,* 4(2), pp. 231-261.

Eng, D. L. with Halberstam, J. and Munoz, J.E. (2005) What's Queer about Queer Studies Now? *Social Text,* 23(3-4), pp. 84-85.

Ensmenger, N (2010). *The computer boys take over: computers, programmers, and the politics of technical expertise.* London: MIT Press.

Epstein, S (2007). *Inclusion: the politics of difference in medical research*. Chicago: University of Chicago Press.

Erdélyi, O. and Goldsmith, J. (2018). Regulating Artificial Intelligence: Proposal for a Global Solution. In: *AIES '18 Proceedings of the 2018 AAAI/ACM, NEW YORK NY, USA. Conference on AI, Ethics, and Society.* [online] Available at: http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_13.pdf

Esteva, A. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature,* 542, pp.115-118.

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

European Commission. Directorate-General for Research Innovation, (2013). *Gendered innovations: How gender analysis contributes to research: Report of the expert group 'Innovation through gender'* (EUR (Luxembourg), 25848). Luxembourg: Publications Office.

European Group on Ethics in Science and New Technologies, (2018). *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems'.* [online] Available at: https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf

Executive Office of the President, (2016). *Artificial Intelligence, Automation, and the Economy.* [online] US Government. Available at: https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF

Executive Order No. 13,859, 84 Fed. Reg. 3967, (February 14, 2019). *Maintaining American Leadership in Artificial Intelligence.* [online] Available at: https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

Feldman, M. et al. (2015). Certifying and Removing Disparate Impact. [online] *Arxiv.* [Preprint] Available at: https://arxiv.org/abs/1412.3756

Ferrando, F. (2014). Is the post-human a post-woman? Cyborgs, robots, artificial intelligence and the futures of gender: A case study. *European Journal of Futures Research,* 2(1), pp.1-17.

Ford, H. and Wajcman, J. (2017). 'Anyone can edit', not everyone does: Wikipedia's infrastructure and the gender gap. *Social Studies of Science,* 47(4), pp.511-527.

Form, M. (2015). *Rise of the robots: technology and the threat of a jobless future. New York:* Basic Books.

Foucault, M. (1991 [1975]). *Discipline and punish: The birth of the prison.* London: Penguin Books.

Freeman, E. (2011). *Time Binds: Queer Temporalities, Queer Histories.* North Carolina: Duke University Press.

Gandy, O (1993). *The Panoptic Sort: A Political Economy Of Personal Information.* Westview Press.

Gendered Innovations n.d., *Case Study: Machine Translation: Analyzing Gender.* [online] Available at: http://genderedinnovations.stanford.edu/case-studies/nlp.html#tabs-2

Gitelman, Lisa (2013). *Raw data is an oxymoron.* Cambridge, Mass: The MIT Press.

Gokul, B (2018). Artificial Intelligence in Financial Services. *Sansmaran Research Journal,* 8(1), pp.3-5.

Goldstaub, T. (forthcoming 2019) *How to talk to Robots.* London: Fourth Estate.

Graells-Garrido, E., Lalmas, M. and Menczer, F. (2011). First Women, Second Sex: Gender Bias in Wikipedia. In: *HT '15 Proceedings of the 26th ACM, NEW YORK NY, USA. Conference on Hypertext & Social Media.* [online] Guzelyurt, Northern Cyprus: ACM, New York NY, US, pp. 165-174. Available at: http://dx.doi.org/10.1145/2700171.2791036

Grgić-Hlača, N. et al (2018). Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In: *WWW 2018: The 2018 Web Conference, April 23–27, 2018.* [online] Lyon, France: ACM, NEW YORK NY, USA. , pp. 903-912. Available at: https://doi.org/10.1145/3178876.3186138

Grgić-Hlača, N. et al (2018). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In: *The Association for the Advancement of Artificial Intelligence conference (AAAI), 2018* [online] New Orleans, USA: AAAI. Available at: https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16523

Grosz, E. (1994). *Volatile bodies: Toward a corporeal feminism.* Bloomington: Indiana University Press.

Halberstam, J. (1991). Automating Gender: Postmodern Feminism in the Age of the Intelligent Machine. *Feminist Studies,* [online] 17(3), pp. 439–460.Available at: JSTOR, www.jstor.org/stable/3178281

Hall, W. and Pesenti, J. (2017). *Growing the Artificial Intelligence Industry in the UK.* [online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf

Hamidi, F. et al. (2018). Gender Recognitions or Gender Reductionism? The Social Implication of Automatic Gender Recognition Systems. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Paper No. 8.* Montreal: AML, pp. 1-13.

Hamlyn, R., Matthews, P. and Shanahan, M. (2017). *Science Education Tracker: Young people's awareness and attitudes towards machine learning.* [online] Kantar Public. Available at: https://wellcome.ac.uk/sites/default/files/science-education-tracker-report-feb17.pdf

Harari, Y. (2017). Reboot for the AI revolution. *Nature,* 550, pp.324-327.

Haraway, D. (1985). Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980s. *Socialist Review,* 80, pp.65-108.

Haraway, D. (1997). Modest_Witness@Second_Millenium.FemaleMan[C]_Meets_OncoMou seTM: Feminism and technoscience. *Journal of the History of Biology,* 30(3) pp. 494-497.

Harding, S. (2009). Postcolonial and feminist philosophies of science and technology: convergences and dissonances. *Postcolonial Studies,* 12:4, pp.401-421.

Harding, S. (2010). *Standpoint methodologies and epistemologies: a logic of scientific inquiry for people.* UNESCO and International Social Science Council, 2010, World Social Science Report, pp.173-175.

Harris, A. (1990). Race and Essentialism in Feminist Legal Theory. *Stanford Law Review,* 42(3), pp. 581-616.

Hawkesworth, J. and Berriman, R. (2018). *Will robots really steal our jobs? An international analysis of the potential long term impact of automation.* [online] PwC. Available at: https://www.pwc.co.uk/economic-services/assets/international-impact-of-automation-feb-2018.pdf

Hawkesworth, M. (1994). Policy studies within a feminist frame. *Policy Sciences,* 27, pp.97-118.

Heilman, M. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior,* 32, pp.113-135.

Hicks, M. (2013). De-Programming the History of Computing. *IEEE Annuals of the History of Computing,* 35(1), pp.86-88.

Hicks, M. (2015). Using Digital Tools for Classroom Activism: Exploring Gender, Infrastructure, and Technological Discipline through a Public Bathroom Project. *Syllabus,* 4(2), pp.1-5.

Hicks, M. (2018). *Programmed inequality: How Britain discarded women technologists and lost its edge in computing.* Cambridge, MA, London, UK: MIT Press.

Hird, M and Roberts, C (2011). Feminism theorises the nonhuman. *Feminist Theory,* 12(2), pp.109-117.

hooks, b. (1996). *Reel to Real: Race, sex, and class at the movies.* New York, London: Routledge.

Information Commissioner's Office, Dipple-Johnstone, J. (2018). *ICO Statement in Response to Facebook Data Breach Announcement.* Available at: https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2018/09/ico-statement-in-response-to-facebook-data-breach-announcement/

Jamnik, M. (2005). *Computer scientist and a woman?* [online] Available at: https://www.cl.cam.ac.uk/~mj201/publications/Computer_Scientist_and_a_Woman.pdf

Johnson, J. (2019). Gender pay gap is getting worse in nearly half of firms, analysis suggests, as critics say forcing firms to report is not enough. *The Telegraph.* [online] Available at: https://www.telegraph.co.uk/news/2019/02/20/gender-pay-gap-getting-worse-nearly-half-firms-analysis-suggests/

Jue Li, J., Ju, W. and Reeves, B. (2017). Touching a mechanical body: tactile contact with body parts of a humanoid robot is psychologically arousing. *Journal of Human-Robot Interaction,* 6(3), pp.118-130.

Kember, S. (2016). *iMedia. The gendering of objects, environments and smart materials.* UK: Palgrave MACM, New York NY, USA. illan.

Keyes. O, (2018). The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. In *Proceedings of the ACM, New York NY, US. on Human-Computer Interaction - CSCW archive.* 2 (CSCW) Article 88. [online] 22 pages Available at: https://doi.org/10.1145/3274357

Keyes, O., Hoy, J. and Drouhard, M. (2019). Human- Computer Insurrection: Notes on an Anarchist HCI. *In CHI Conference on Human Factors in Computing Systems Proceedings* (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK. ACM, New York, USA. [online] 13 pages. Available at: https://doi.org/10.1145/3290605.3300569

Khullar, D. (2019). A.I. could worse Health Disparities. *New York Times.* [online] Available at: https://www.nytimes.com/2019/01/31/opinion/ai-bias-healthcare.html

Kilbertus, N. et al. (2018). Blind Justice: Fairness with Encrypted Sensitive Attributes. In: *Proceedings of the 35th International Conference on Machine Learning.* Stockholm. Sweden: PMLR, 80, pp. 2630-2639.

Kim, M., Ghorbani, A. and Zou, J. (2018). Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. [online] *Arxiv.* [Preprint] Available at: https://arxiv.org/abs/1805.12317

Kusner, M. et al. (2017). Counterfactual Fairness. 31st Conference on Neural Information Processing Systems (NIPS 2017). [online] *Arxiv.* [Preprint] Available at: https://arxiv.org/pdf/1703.06856.pdf

Laboratory for Computer Science and the Artificial Intelligence Laboratory at MIT: Prepared by female graduate students and research staff, (1983). *Barriers to Equality in Academia: Women in Computer Science at MIT* [online] Available at: https://homes.cs.washington.edu/~lazowska/mit/

Lambrecht, A. and Tucker, C.E. (2018). *Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads.* [online] Available at SSRN: https://ssrn.com/abstract=2852260 or http://dx.doi.org/10.2139/ssrn.2852260

Leavy, S. (2018). Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. In: 1st International Workshop on Gender Equality in Software Engineering (GE). New York NY, US: ACM, pp.14-16.

Leavy, S. (2019). Uncovering gender bias in newspaper coverage of Irish politicians using machine learning. *Digital Scholarship in the Humanities,* 34(1), pp.48-63.

Levit, N. and Verchick, R. (2006). *Feminist legal theory: A primer.* New York: New York University Press.

Lewthwaite, S. and Jamieson, L. (2019). *Big Qual – Why we should be thinking big about qualitative data for research, teaching and policy.* [Blog] London School of Economics, The Impact Blog. Available at: https://blogs.lse.ac.uk/impactofsocialsciences/2019/03/04/big-qual-why-we-should-be-thinking-big-about-qualitative-data-for-research-teaching-and-policy/

Liberty – Written evidence (AIC0181) (2017) [online] Available at: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69699.html

Light, Jennifer (1999). 'When Computers Were Women'. *Technology and Culture,* 40(3), pp.455-483.

Longino, H. and Lennon, K. (1997). Feminist Epistemology as a Local Epistemology. *Proceedings of the Aristotelian Society,* 71, pp.19-35; 37-54.

Louizos, C. et al. (2016). The Variational Vair Autoencoder. [online] *Arxiv.* [Preprint] Available at: https://arxiv.org/abs/1511.00830

Lyness, K., Heilman, M., and Zedeck, S. (2006). When Fit is Fundamental: Performance Evaluations and Promotions of Upper-Level Female and Male Managers. *Journal of Applied Psychology,* 91(4), pp.777-785.

Mahendran, D D (2011). Race and Computation: An Existential Phenomenological Inquiry Concerning Man, Mind, and the Body. Unpublished doctoral thesis. Berkley: University of California.

Mantha, Y. and Hudson, S. (2018). *Estimating the Gender Ratio of AI Researchers Around the World.* [online] Element AI Lab. Available at: https://medium.com/element-ai-research-lab/estimating-the-gender-ratio-of-ai-researchers-around-the-world-81d2b8dbe9c3

Martell, R.F. et al. (1998). Sex stereotyping in the executive suite: 'Much ado about something'. *Journal of Social Behavior and Personality,* 13, pp.127–138.

Mateescu, A. and Elish, M. C. (2019). *AI in Context: The Labor of Integrating New Technologies.* [online] Data & Society. Available at: https://datasociety.net/wp-content/uploads/2019/01/DataandSociety_AIinContext.pdf

McAfee, A. and Brynjolfsson, E. (2016). Human Work in the Robotic Future: Policy for the Age of Automation. [online] *Foreign Affairs.* Available at: https://www.foreignaffairs.com/articles/2016-06-13/human-work-robotic-future

McQuillan, Dan (2019). Towards an anti-fascist AI. Zenodo. Available from: http://doi.org/10.5281/zenodo.2649824'

Misa, T (2010). *Gender codes: why women are leaving computing.* Hoboken: Wiley.

Mishra, V. and Srikumar, M. (2017). Predatory Data: Gender Bias in Artificial Intelligence. [online] In: Saran, Samir (2017). *Digital Debates. CyFy Journal.* Available at: https://www.orfonline.org/wp-content/uploads/2017/10/CyFy_2017_Journal.pdf

Niklas, Jedrzej (2016). E-government in the welfare state – human rights implications of digitization of social policy in Poland. *Global Information Society Watch,* 10th Edition, pp.182-188.

Ní Loideáin, N. and Adams, R. (2018) *From Alexa to Siri and the GDPR: The Gendering of Virtual Personal Assistants and the Role of EU Data Protection Law.* [online] King's College London Dickson Poon School of Law Legal Studies Research Paper Series. Available at SSRN: https://ssrn.com/abstract=3281807 or http://dx.doi.org/10.2139/ssrn.3281807

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

Nussbaum, M. (2011). *Creating Capabilities: the human development approach.* Cambridge: Harvard University Press.

O'Connor, S. (2019). The robot-proof skills that give women an edge in the age of AI. [online] *Financial Times.* Available at: https://www.ft.com/content/06afd24a-2dfb-11e9-ba00-0251022932c8

O'Neill, C. (2016). *Weapons of Math Destruction.* New York: Crown Books.

Orwell, G. (1987 [1949]). *Nineteen Eighty-Four.* London: Penguin.

Parliament. House of Lords Artificial Intelligence Select Committee. (2017). *Report on 'AI in the UK: Ready, Willing and Able?* [online] Report of Session 2017-2019, HL Paper 100 (April 2017). Available at: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

Parliament. House of Lords Select Committee on Artificial Intelligence. (2017). Collated Written Evidence Volume. Available at: https://www.parliament.uk/documents/lords-committees/Artificial-Intelligence/AI-Written-Evidence-Volume.pdf

Oyěwùmí, O. (1997) *Invention of Women: Making an African Sense of Western Gender Discourses.* Minneapolis; London: University of Minnesota Press.

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

Popejoy, A and Fullerton, S. (2016). Genomics is failing on diversity. *Nature,* 538 (7624), pp.161-164.

Powell, G., Butterfield, A.and Parent, J. (2002). Gender and Managerial Stereotypes: Have the Times Changed? *Journal of Management,* 28(2), pp.177-193.

Puar, J (2007). *Terrorist assemblages.* Durham: Duke University Press.

Radford, A. et al. (2019). Language Models are Unsupervised Multitask Learners. Open AI. [online] Available at: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rankin, J (2018). *A people's history of computing in the United States.* Cambridge, Mass: Harvard University Press.

Rankin, J (2015). Why I love 'The Bletchley Circle' and you should, too. [online] *Lady Science,* Available at: https://www.ladyscience.com/bletchley-circle/ynf2fugzghrv6atla7plfnimf2iu08

Rankin, J (2015). Queens of Code. [online] *Lady Science,* Available at: https://www.ladyscience.com/queens-of-code

Ritschel, C. (2018). Google Fixes Translate Tool to Correct Gendered Pronouns. *The Independent.* [online] Available at: https://www.independent.co.uk/life-style/women/google-translate-sexist-masculine-feminine-he-said-she-said-english-spanish-languages-a8672586.html

Rousseau, JJ. (1762 [1997]). *'The Social Contract' and Other Later Political Writings, (Cambridge Texts in the History of Political Thought),* V. Gourevitch, ed., Cambridge: Cambridge University Press.

Sanders, J. (2005). Gender and technology in education: What the research tells us. In: *Proceedings of the International Symposium on Women and ICT,* 126.

Schiebinger, L. and Schraudner, M. (2011). Interdisciplinary Approaches to Achieving Gendered Innovations in Science, Medicine, and Engineering. *Interdisciplinary Science Review,* 36, no.2, pp. 154-167.

Schools, pupils and their characteristics (January 2018). Department for Education. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/719226/Schools_Pupils_and_their_Characteristics_2018_Main_Text.pdf

Shaikh, S. et al. (2017). An End-To-End Machine Learning Pipeline That Ensures Fairness Policies. In: Bloomberg Data for Good Exchange Conference. [online] *Arxiv.* [Preprint] Available at: https://arxiv.org/abs/1710.06876

Shams, Z et al. (2018). Accessible reasoning with diagrams: From cognition to automation. [online] *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 10871 LNAI 247-263. https://doi.org/10.1007/978-3-319-91376-6_25

Shapiro, E (2010). *Gender circuits: bodies and identities in a technological age.* New York: Routledge.

Singh, V. (2002). *Managing Diversity for Strategic Advantage.* London: Council for Excellence in Management and Leadership.

Smith, B. (2018). 'Facial recognition technology: *The need for public regulation and corporate responsibility*. [Blog] Microsoft Blog. Available at: https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/

Snorton, C. R. (2017). *Black on Both Sides: A Racial History of Trans Identity.* Minneapolis: University of Minnesota Press.

Speicher, T. et al. (2018). A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. In: *KDD '18: SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018,* [online] London, United Kingdom. New York, US: ACM. 12 pages. Available at: https://doi.org/10.1145/3219819.3220046

Spiel, K., Keyes, O. and Barlas, P. (2019). Patching Gender: Non-binary Utopias in HCI. In: *CHI EA '19 Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.* [online] Glasgow, Scotland. ACM, New York NY, USA. , New York, USA. Available at: https://doi.org/10.1145/3290607.3310425

Spivak, G. (1988) 'Can the Subaltern Speak?', in Nelson, C. and Grossberg, L. (eds.) *Marxism and the Interpretation of Culture.* Urbana: University of Illinois Press, pp. 271-313.

Spivak, G. (1988) *In Other Worlds: Essays in Cultural Politics.* New York: Routledge.

Stanley, E. and Smith, N. (2011). *Captive genders: Trans embodiment and the prison industrial complex.* Edinburgh: AK Press.

Stone, S. (1992) The Empire Strikes Back: A Postranssexual Manifesto. *Camera Obscura,* 10(2 29), pp. 150-176.

Stryker, S. (2008). *Transgender history. Berkeley*, CA: Seal Press.

Stryker, S. and Whittle, S. (2006). *The Transgender Studies Reader.* New York: Routledge.

Stryker, S. (2004). Transgender Studies: Queer Theory's Evil Twin. GLQ: *A Journal of Lesbian and Gay Studies,* 10(2), pp.212-215.

The Centre for Data Ethics and Innovation: Review on Bias in Algorithmic Decision Making (2019). [online] Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/799646/CDEI_Bias_Review_ToR_and_Call_for_Evidence_2019.pdf

The European Commission's High-Level Expert Group on Artificial Intelligence (2018). Draft Ethics Guidelines for Trustworthy AI: Executive Summary.

Thompson, C. (2019). The Secret History of Women in Coding. *New York Times.* [online] Available at: https://www.nytimes.com/2019/02/13/magazine/women-coding-computer-programming.html

Turing, A. (1950). Computing Machinery and Intelligence. *Mind,* 59(236), pp.433-460.

Vaidhyanathan, S (2011). *The Googlization of everything: (and why we should worry).* Berkeley: University of California Press.

Varley, C. (2018). Are sex robots just turning women into literal objects? *BBC News.* [online] Available at: https://www.bbc.co.uk/bbcthree/article/8bbe0749-62ee-40f9-a8ac-a2d751c474f6

Villani, C. (2018). *For a Meaningful Artificial Intelligence: Towards a French and European Strategy. Mission assigned by the Prime Minister Édouard Philippe* [online] Available at: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

Vincent, J. (2018). Google Removes Gendered Pronouns from Gmail's Smart Compose to Avoid AI Bias. *The Verge.* [online] Available at: https://www.theverge.com/2018/11/27/18114127/google-gmail-smart-compose-ai-gender-bias-prounouns-removed

Wajcman, J. (2006) The Feminisation of Work in the Information Age? In: M. Frank Fox, D. Johnson and S. Rosser, eds., *Women, Gender, and Technology.* Champaign, Ill.: University of Illinois Press, pp. 80-97.

Wajcman, J. (2006). Suspending Gender? Reflecting on Innovations in Cyberspace. In: H. Nowotny, ed., *Cultures of Technology and the Quest for Innovation.* New York: Berhahn Books, 2006, pp. 95-110.

Wajcman, J. (2007) From Women and Technology to Gendered Technoscience. *Information, Community and Society,* [online] 10(3), pp.287-298. Available at: doi: 10.1080/13691180701409770

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

Ware, S. M. (2017). All Power to All People? Black LGBTTI2QQ Activism, Remembrance, and Archiving in Toronto. *TSQ,* 4 (2), pp.170–180.

Warner, M. (1999) *The Trouble with Normal: Sex, politics and the ethics of queer life.* Cambridge, MA: Harvard University Press.

Weber S. and KRC Research. (2016). *AI-Ready or Not: Artificial Intelligence Here We Come! What Consumers Think & What Marketers Need to Know.* New York: Weber Shandwick.

West, M., Kraut, R., and Chew, H. E., (2019). I'd blush if I could: closing gender divides in digital skills through education. *UNESCO and EQUALS Skills Coalition.* Available from: https://unesdoc.unesco.org/in/documentViewer.xhtml?v=2.1.196&id=p::usmarcdef_0000367416&-file=/in/rest/annotationSVC/DownloadWatermarkedAttachment/attach_import_77988d38-b8bd-4cc1-b9b4-3cc16d631bf9%3F_%3D367416eng.pdf&locale=en&multi=true&ark=/ark:/48223/pf0000367416/PDF/367416eng.pdf#%5B%7B%22num%22%3A384%2C%22gen%22%3A0%7D%2C%7B%22name%22%3A%22XYZ%22%7D%2C0%2C842%2Cnull%5D

White, M. (2015). *Producing women: The Internet, traditional femininity, queerness, and creativity.* London: Routledge.

Whittaker, M. et al. (2018). *AI Now Report 2018.* [online] Available at: http://blog.quintarelli.it/wp-content/blogs.dir/10121/files/2018/12/AI_Now_2018_Report.pdf

Whittlestone, J. et al. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* London: Nuffield Foundation.

Wilcox, L. (2017). Embodying algorithmic war: gender, race, and the posthuman in drone warfare. *Security Dialogue,* 48(1), pp. 11–28.

Wille, B. et al. (2018). Personality characteristics of male and female executives: Distinct pathways to success? *Journal of Vocational Behavior,* 106.

Willemsen, T. (2002). Gender Typing of the Successful Manager: A Stereotype Reconsidered. *Sex Roles,* 46(11), pp.385-391.

Winfield A. and Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Phil. Trans. R. Soc. A* [online] 376: 20180085. Available at: http://dx.doi.org/10.1098/rsta.2018.0085

Woods, H. S. (2018). Asking more of Siri and Alexa: Feminine persona in service of surveillance capitalism. *Critical Studies in Media Communication,* 35(4), pp. 334-349.

World Economic Forum (2018). *Agile governance: reimagining policy-making in the fourth industrial revolution.* [White paper] [online] Available at: http://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf

Zafar, M. B. et al. (2017). Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of Machine Learning Research* [online] PMLR, 54, pp 962-970.

Zemel, R. et al. (2013). Learning Fair Representations. In: *ICML'13 Proceedings of the 30th International Conference on Machine Learning.* New York, NY, US: ACM, 28, III-325- 333.

Zhao, J. et al. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. Empirical Methods in Natural Language Processing. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: ACL, pp. 2979—2989.

Zook, M. et al. (2017). Ten simple rules for responsible big data research. *PLoS Computational Biology,* [online] 13 (3) Available at: e1005399. doi:10.1371/journal.pcbi.1005399. http:// dx.doi.org/10.1371/journal.pcbi.1005399

Zou, J. and Schiebinger, L. (2018). AI can be sexist and racist – it's time to make it fair. *Nature.* [online] 559(7714), pp. 324-326. Available at: https://www.nature.com/articles/d41586-018-05707-8

Appendix 1:

List of Groups for Collective Intelligence Activity

Group 1 The Gender Binary: Epistemological, Physiological and Linguistic Gender Stereotypes in AI	Group 2 The Gender Politics of AI: Ethics, Policy and Privacy	Group 3 Data, Discrimination and Diversity	Group 4 AI and Gender in Organisations
Sarah Dillon	Steven Cave	Jess Whittlestone	Jonnie Penn
Lauren Wilcox	Jude Browne	Joy Lisi Rankin	Rob McCargow
Os Keyes	Nóra Ni Loideain	Susan Leavy	Rumman Chowdhury
Alison Adam	Rachel Adams	Adrian Weller	Gina Neff
Judy Wacjman	Julian Huppert	Londa Schiebinger	Elena Sinel
Beth Singler	Mateja Jamnik	Theo Bearman	Peter He
Kanta Dihal	Christopher Markou	Allison Gardner	Maria Axente
Gabrielle McGuinness	Ivana Bartoletti	Jennifer Cobbe	Tim Gardam
Isabelle Guenette Thornton	Zohreh Shams	Olivia Varley-Winter	Diana Robinson
Seab	Anna Alexandrova	Elena Rastorgueva	
Toshie Takahashi	Brittany Smith	Ezinne Nwankwo	
	Reema Patel	Rafael Calvo	
	Tabitha Goldstaub		

Appendix 2:

List of Attendees

Alison Adam
Rachel Adams
Anna Alexandrova
Maria Axente
Ivana Bartoletti
Theo Bearman
Jude Browne
Rafael Calvo
Stephen Cave
Rumman Chowdhury
Jennifer Cobbe
Clementine Collett
Kanta Dihal
Sarah Dillon
Tim Gardam
Allison Gardner
Tabitha Goldstaub
Peter He
Julian Huppert
Mateja Jamnik
Os Keyes
Susan Leavy
Nóra Ni Loideain
Christopher Markou
Rob McCargow
Gabrielle McGuinness
Gaenor Moore
Gina Neff
Ezinne Nwankwo
Reema Patel
Jonnie Penn
Joy Lisi Rankin
Elena Rastorgueva
Diana Robinson
Londa Schiebinger
Zohreh Shams
Elena Sinel
Beth Singler
Brittany Smith
Toshie Takahashi
Isabelle Guenette Thornton
Judy Wacjman
Adrian Weller
Jess Whittlestone
Lauren Wilcox
Lucy van de Wiel
Olivia Varley-Winter

Appendix 3:

List of Speakers and Abstracts

Panel 1 – History, Narrative, Theory: Interdisciplinary Perspectives

Dr Joy Lisi Rankin, Author of A People's History of Computing in the United States.
Title: Whose Intelligence(s)?: A research agenda inspired by 1960s American educational computing

Abstract: In 1960 engineers at the University of Illinois began exploring the uses of computing in education; a decade later, they were expanding their networked system to include hundreds of plasma-screen terminals around the United States. Thousands of PLATO people performed gender on and through the system in multifarious ways.

This paper examines the nursing lessons created for the system by Maryann Bitzer; her students learned about heart attacks and prenatal care via PLATO courses. I analyze how Bitzer created these lessons, including subject matter, and how students experienced this acquisition of intelligence to suggest a research agenda encompassing AI, gender, and race. Bitzer emphasized the technologies – of computing, of education – driving her work; however, my focus on users, and the people (and assumptions) populating Bitzer's lessons, yields questions: AI by, for, and about whom?

I underscore the methodological value of PLATO's origins in education, with project publications explicitly describing teaching and learning, including how users encountered terminals, computing languages, and similar issues. PLATO researchers presented a model of the computer teaching the student (under the guise of students directing their own learning). Yet, PLATO as a technology masked the fact that people were always behind the scenes, creating those lessons (and the sociotechnical systems in which they were embedded). In other words, people "taught" the computer teaching the student. Thus, the educational context helps illuminate contemporary research about AI by drawing out the invisible, the hidden, and the assumed about human and machine learning, human and machine intelligence – and directing our research there.

Professor Alison Adam, Professor of Science, Technology and Society, Sheffield Hallam University, UK.
Title: Reflecting on the history of gender and AI

Abstract: As we attempt to map out a research agenda for gender and AI we should consider how the history of gender and AI may offer useful insights. When I first started researching gender and AI some twenty-five or so years ago it appeared to be a somewhat niche area, even though there was plenty of good research on gender and science and technology in the UK and elsewhere. My research on gender and AI, at that time, centred on feminist critiques of philosophical approaches to AI and attempts to uncover the ways in which existing AI-based systems reflected masculine approaches to knowledge. Such systems were largely research tools, game-playing systems, artificial societies, situated robotics and the like. None of those systems was used commercially at the time and it is this aspect which marks the difference between a 'then' of 20+ years ago and now where AI systems have begun to proliferate. As we continue to consider how gender, sexuality, class and race are inscribed in the design of technological systems, recent examples, such as the Amazon 'sexist AI recruitment tool', demonstrate that we are far from achieving equality. While earlier feminist critiques of AI initially focused on ways in which AI was gendered male and how it ignored woman's ways of knowing in the design of technological systems, we now have the spectre of AI systems which are consciously modelled on a potentially disturbing model of femininity as digital handmaids - intelligent assistants such as Alexa, Siri and Tay.ai.

Dr Sarah Dillon, Programme Director, AI: Narratives and Justice, CFI; University Lecturer in Literature and Film, Faculty of English, University of Cambridge.
Title: The Societal Harm of Gendering VPAs: Reasoning by Analogy with Eliza and Pygmalion

Abstract: This paper presents the key insights of a forthcoming article which synthesises popular media arguments regarding the reasons for, and consequences of, the gendering of Virtual Personal Assistants (VPAs) and identifies emerging academic scholarship in the field. The article then proposes that methodologies from the humanities, including from history, philosophy and literary studies, can be used to expand the evidence base that such gendering causes societal harm. This is part of broader case that attention to literature and other forms of fictional narrative must be included in the sociological study of scientific knowledge, and that humanities methodologies must be included in the study of social effects of emerging technologies. The articles takes as its case study Joshep Weizembaum's natural language processing programme, ELIZA, which he named after the character in George Bernard Shaw's Pygmalion. The article introduces and employs two methodologies to demonstrate how this case study informs contemporary debate: close reading, and reasoning by analogy. The article proposes that both

are key techniques of feminist critique. These techniques are employed in the article in order to debate regarding VPAs and societal harm: the relationship between the natural and the artificial; the objectification of women; the gendered power and subservience.

Dr Lauren Wilcox, Deputy Director of the University of Cambridge Centre for Gender Studies, UK.
Title: Robots and Avatars: Gender and Race in AI Imaginaries

Abstract: This talk draws particular from a chapter my work in progress 'War Beyond the Human' that centres the figuration of the robot, which serves as a nexus of many intertwined imaginaries and materializations of artificially intelligent machines. The 'robot' serves as an avatar that represents the human: standing in for, but at the same time displacing the human. To understand the critical potentials of thinking through relations of violence, embodiment, race and capitalism that are opened up by this figuration, we first need to interrogate the concept of 'gender' to understand the limits of contemporary feminist critiques of the figure of the 'robot' in culture and society. I will provide a brief genealogy of the concept of gender and of the effects that 'gender' may be said to have. First, gender is always already 'queer': that is, gender is about the regulation of sexuality. Second, gender is technological. Third, gender is a racializing apparatus. My intervention in this work is to argue that not only is AI 'gendered' but also that our understanding of 'gender' is rooted in a similar epistemic space. Recognizing these connections helps us to understand how questions of race and racialization are often elided in critical discourses of AI.

Panel 2 – Trust, Transparency and Regulation

Dr Tameem Adel, Machine Learning Group, University of Cambridge, UK, CFI Research Fellow on 'Trust and Transparency'.
Title: Current Technical Work on Fairness in Machine Learning

Dr Nóra Ni Loideain, Lecturer in Law and Director of Information Law and Policy Centre, Institute of Advanced Legal Studies, School of Advanced Study, University of London, UK.
Title: From Ava to Siri: Gendering VPAs and the Role of EU Data Protection Law

Abstract: With female names, voices and characters, artificially intelligent Virtual Personal Assistants such as Alexa, Cortana, and Siri appear to be decisively gendered female. Through an exploration of the various facets of gendering at play in the design of Siri, Alexa and Cortana, we argue that this gendering of VPAs as female may pose a societal harm, insofar as they reproduce normative assumptions about the role of women as submissive and secondary to men. In response, this paper examines the potential role and scope of data protection law as one possible solution to this problem. In particular, we examine the role of data privacy impact assessments that highlight the need to go beyond the data privacy paradigm, and require data controllers to consider and address the social impact of their products

Professor Londa Shiebinger. John L. Hinds Professor of History of Science, History Department, Stanford University, US. Director of the EU/US Gendered Innovations in Science, Health & Medicine, Engineering, and Environment Project. Director of Stanford's Clayman Institute for Gender Research (2004-2010).
Title: The Future of Human-Centered AI? Open Questions and Collaborations.

Abstract: Stanford is announcing a Human-Centered AI Institute in March. Does "human-centered" make for socially-responsible AI? What configurations of researchers could make that happen? What is required for fruitful collaborations between humanists and technologists?

Dr Adrian Weller, Machine Learning, University of Cambridge, UK. Programme Director for AI at The Alan Turing Institute and Board Member of the Centre for Data Ethics and Innovation, UK.
Title: Why We Should Care About Diversity in AI Research.

Ms Elena Sinel, Acorn Aspirations and Teens in AI
Title: Teens in AI

Abstract: Teens in AI is Acorn Aspirations' special initiative launched at the UN AI for Good Global Summit to democratise AI and create pipelines for underrepresented talent, thereby improving diversity and inclusion in Artificial Intelligence. We offer young people aged 12-18 early exposure to AI for social good through a combination of expert mentoring, talks, workshops in AI/ML, human-centred design and ethics, hackathons, accelerators, company tours and networking opportunities. Elena Sinel and Peter He, co-founders of Teens in AI, will present experiences, findings, insights and some open questions from their work.

Panel 3 – Organisational Initiatives to Increase Gender Equality

Mr Rob McCargow, Director of AI, PwC UK
Background and work: Director of AI at PwC UK; works to drive innovation within the firm and develop new services for clients. Works for responsible technology and promotes awareness of the growing ethical agenda relating to AI. On the advisory board for the All-Party Parliamentary Group on AI, an adviser to the IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems and a TEDx Speaker.

Abstract: This talk will detail a range of gender diversity initiatives with which PwC has been involved, reflect on their effectiveness to address prevalent problems and challenges in relation to AI and Gender, and speculate on where or what more could be done.

Dr Mateja Jamnik, Reader in Artificial Intelligence, Department of Computer Science and Technology, University of Cambridge, UK.
Title: Developing AI Technology Without Bias?

Abstract: AI systems are becoming a ubiquitous part of our lives. It is therefore paramount that when AI technology is used to make decisions about human lives, this is done in a fair, unbiased and transparent way. Current policies, such as GDPR are trying to legislate this basic right. Unfortunately, there are many ways in which AI technology cannot and does not fulfil this right, in particular for minority gender and ethnic groups. In terms of bias, there are at least two sources where bias in AI technology originates from. First is the AI (machine learning) algorithms: they are built by developer groups that lack diversity, so their design decisions reflect their (unrepresentative) view of the world. Second is the data that the AI algorithms are learning from: it reflects our biased society where minority groups do not have a fair representative access to generating data. We know that what is good for diverse groups is good for everybody, so how can we mitigate against biases and ensure fair decision making by AI technology?

Dr Susan Leavy, Post-Doc at Insight Centre for Data Analytics, University College Dublin, Ireland.
Title: Algorithmic Bias: Gender Proofing AI

Abstract: The increasingly widespread use of Artificial Intelligence has the potential to set back decades of advances in gender equality in society. Gender bias has been uncovered in AI systems promoting job advertisements, facial recognition systems and neural word embedding models used in web search and recommender systems. Machine learning algorithms reflect the kinds of bias inherent in training data and there is a growing awareness of the dangers posed by the potential of AI to reinforce societal biases. Work is ongoing to prevent this through testing of learned associations, introduction of concepts of fairness to machine learning algorithms and analysis of training data. However preventing the learning of gender bias can be particularly problematic when algorithms are trained on language based corpora. There is a wealth of scholarship in the areas of gender theory, feminist linguistics and sociolinguistics that deconstructs how gender ideology is embedded in language and the way it is used in society. Integrating this work with AI is key to understanding how gender bias is learned and how it may be removed from training corpora. This paper demonstrates how bridging AI and research in gender and language can provide a framework for gender proofing AI and enable the systematic detection and prevention of algorithmic gender bias.

Professor Gina Neff, Senior Research Fellow and Associate Professor at the Oxford Internet Institute and at the Department of Sociology, University of Oxford. Faculty affiliate of the Center on Organizational Innovation, Columbia University, USA.
Title: Automating gendered identity? The challenges for implementation and management of AI in organisations.

Abstract: Will new AI systems exacerbate gender inequality in established workplaces and organizations? This talk presents historical research on the efforts to digitize health care and construction to suggest concrete ways the widespread adoption of AI systems will worsen gender inequality. The single biggest challenge for ethical AI will be how companies deploy, adopt and adapt to AI systems and that their outcomes will be shaped as much by social factors as by technical ones. I will look at three key sociological concepts that shape whether AI systems are successfully used in workplaces: organizational hierarchy, organizational routines, and institutional power and how these intersect with gender dynamics. Then, using emerging use and abuse cases, I develop a framework for managing in AI-enabled workplaces that begins to build 'human-centred' AI.

Panel 4 – Challenging Built in Bias and Gender Stereotypes

Dr Rachel Adams, Information, Law and Policy Centre, Institute of Advanced Legal Studies, School of Advanced Study, University of London, UK.
Title: "Make Google do it": Interpellating digital subject/objects of desire in the gendering of artificially intelligent virtual personal assistants

Abstract: Most of the artificially intelligent virtual personal assistants on the market today are gendered female, with female names and characters - evident both in their design and marketing. But this gendering - which has been noted in scholarship - meets with an imperative grammar which users are given to engage with them, most notably the call "Hey Siri!" or "Hey Alexa!" Drawing on Louis Althusser and Judith Butler's accounts of interpellation, I discuss how the gendering of these devices coupled with the new techno-crafted grammar of the interaction with their users, plays into a broader history of female automata produced as both subjects to and objects of a male desire to (literally) constitute, control and command the female. Thus, expanding on the current account of female gendered virtual personal assistants as a strategy for softening the transition from a traditional past to a disruptive digital future, I argue that, more critically, this gendering re-inscribes and legitimises a historically embedded male desire which works, primarily, to ensure the female is always in a subject to/object of relationship to their male "user"/maker, without agency to attain her own status as subject.

Dr Rumman Chowdhury, Global Lead for Applied AI, Accenture Applied Intelligence.
Title: Algorithmic Determinism: Digital reinforcement of gender norms

Abstract: We have evolved our use of analytic technologies from retrospective to predictive to prescriptive. In this talk I introduce Algorithmic Determinism, the generalized notion that probabilistic systems can create self-fulfilling prophecies for predictive subjects which result in a wide range of detrimental effects. This effect arises when accompanied by three factors: assumptions about variable mutability (i.e., the ability for an individual to change an impactful feature) and the combination of measurement bias and feedback loops. In particular, they can serve to reinforce gender norms in the same manner of socialization - subtle incentives to act a particular way that can result in reward or attention. They can also serve to reinforce gender stereotypes, by limiting options or choices due to assumptions about the gender presented.

Mx Os Keyes, Data Ecologies Laboratory, Department of Human Centred Design & Engineering, University of Washington, US. Specialist in Gender Studies, Human Computer Interaction and STS.
Title: Trans Models, Trans Selves: AI and Reinscription of Gender

Abstract: Technologies have long been observed reinscribing normative and constraining models of gender. AI is no different. From gender recognition systems (which assign gender from physiological structure or vocal tones) to algorithms that infer gender from behavioural traces, providing gendered cues in response, the ecosystem is replete with examples of AIs that not only use gender as input, but in doing so, reproduce it.

A particular concern is those systems which reinforce the norm of gender as a physiological attribute, and their current and potential impact on transgender (trans) people. In this talk I will briefly explore some existing work that examines AI's dependence on physiological models of gender, highlighting areas where questions have been left unanswered, before setting out a possible research agenda for the future – one focused on interdisciplinary and grounded work that engages with and defers to trans communities.

Professor Judy Wajcman, Anthony Giddens Professor of Sociology, Department of Sociology, The London School of Economics and Political Science, UK.
Title: Mind the gender gap in expertise: the case of Wikipedia

Abstract: Feminist Science and Technology Studies (STS) has long established that science's provenance as a male domain continues to define what counts as knowledge and expertise. Wikipedia, arguably one of the most powerful sources of information today, was initially lauded as providing the opportunity to rebuild knowledge institutions by providing greater representation of multiple groups. However, less than twenty percent of Wikipedia editors are women. At one level, this imbalance in contributions and therefore content is yet another case of the masculine culture of technoscience. This is an important argument and, in this talk, I will examine the empirical research that highlights these issues. My main objective, however, is to demonstrate that Wikipedia's infrastructure introduces new and less visible sources of gender disparity.


UNIVERSITY OF
CAMBRIDGE


CFI
LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE